Separating the Signal From the Noise: An Examination of Student and Teacher Scores Based on Student Learning Objectives (SLOs) in One State

Separating the Signal from the Noise: An Examination of Student and Teacher Scores Based on Student Learning Objectives (SLOs) in One State

By Katie Hills Buckley

Dissertation Committee:
Heather Hill (Chair)
Andrew Ho
Richard Murnane

A Thesis Presented to the Faculty of the Graduate School of Education of Harvard University in Partial Fulfillment of the Requirements for the Degree of Doctor of Education

2015

Acknowledgement Page

This page was important for me to write, because so many people helped me, both directly and indirectly, complete this dissertation. I owe a huge debt of gratitude to my advisor and dissertation chair, Heather Hill, whose critical feedback, numerous read-throughs, and encouragement of this project helped shape this dissertation and ensure its completion. I wish to also thank my committee member, Andrew Ho, for his tremendous help in shaping the analyses and providing invaluable analytic guidance, and my committee member, Richard Murnane, for his thoughtful feedback and probing questions.

I would also like to thank the state whose data I used for this dissertation. This state and districts in my sample wish to remain anonymous, but to those who helped me obtain permission to use the data, provided the data in a functional format, and addressed each and every one of my questions, I am truly grateful.

I am grateful to have studied under and worked with wonderful teachers, mentors and colleagues. To my high school teachers who encouraged me to think more deeply and creatively; to my college advisor and mentor, Mark Hyde, who shaped my path in higher education through the perfect combination of encouragement and high expectations; and to the amazing professors at HGSE, John Willett and Richard Murnane in particular, whose teaching skills and research guidance are second to none: Thank you! I am fortunate to have worked alongside brilliant colleagues as well, particularly the LT gang, who cheered me on (and at times pulled me along) when I needed it the most. And, I'd certainly be remiss without acknowledging my supervisor at the Center for Assessment, Scott Marion, who took me under his wing, encouraged me to write on SLOs, and has not only informed my thinking on this topic, but has answered countless questions from me along the way. Each of these people helped open doors for me that I didn't believe I could walk through or didn't even realize existed, and I will be forever grateful to them.

I want to acknowledge my family, especially my parents, whose faith in me never faltered and whose generous support of and belief in my education is unsurpassed. Their unconditional love has guided me through many *many* years of higher education. I want to thank my sister, brother-in-law and niece, who were always willing to provide relief – in the form of babysitting and fun trips – whenever I needed it. Thank you to my in-laws, who offered countless home-cooked meals and took care of my daughter during school vacation days and school closings so that I could work on this dissertation.

Last, but certainly not least, I am beyond grateful for the love and support of my husband, Jay. As the Chief Morale Officer of our household, Jay consistently provided the encouragement, perspective, and laughter (not to mention the dark chocolate-covered pretzels and Dunkin Donuts coffee) that I needed to make it through this process. Since I began my doctoral career at Harvard seven years ago, Jay has been by my side and there is no one else I'd rather have as my partner in life.

Finally, I would like to dedicate this dissertation to my daughter, Sadie Grace Higgins. At nearly three years old, she has only a vague understanding that the countless hours her mama spent at the computer were for the purpose of finishing this dissertation. However, it is my hope that through this experience, she comes to understand the incredible value in learning and, most importantly, that she never stops asking “why?”.

Table of Contents

List of Tables and Figures

Abstract

Despite the prevalence of student learning objectives (SLOs) in teacher evaluation systems throughout the United States, research on the validity of student and teacher SLO scores used for high-stakes decisions is lacking. For this reason, this dissertation is comprised of two chapters that examine student and teacher-level SLO performance data from select districts in one Race to the Top state. In Chapter 1, I describe the quality of student assessment data and the comparability of student scores across alternative growth targets. I find that in the first year of implementation, assessments from half of the courses in the sample contained indicators of poor data quality, including anomalous score distributions and small to negative correlations between student prescores and postscores. However, in the second year of implementation, when student SLO performance is incorporated into final teacher evaluation scores, far fewer assessments contained anomalous score distributions, and there is no evidence to suggest manipulation of student scores. In addition to the assessments, the choice of student growth target does have an impact on the comparability of student and teacher scores across districts and years.

Chapter 2 describes the validity and reliability of teacher SLO scores. I find that while teacher SLO scores are moderately stable across courses, they are not stable over time, likely due to changes made to the assessments and targets used to determine student SLO scores. Further, for teachers with both SLO scores and an alternative metric of performance based on student growth, the two metrics do not converge. Finally, teachers in courses with higher average student prescores and lower proportions of students with disabilities have slightly higher SLO scores. In general, results on teacher SLO scores were similar to those found with value-added based metrics of teacher performance. Findings from both chapters suggest that improvement in the quality of the assessments administered as well as greater consistency in the growth targets assigned to students, both within districts over time and across districts, will improve the validity of student and teacher SLO scores in this state.

Dissertation Introduction

Student learning objectives (SLOs), first piloted in Denver Public Schools in 1999 as a process to determine teacher compensation, are now being used in teacher evaluation systems in upwards of 30 states and districts throughout the U.S. SLOs involve a process whereby educators create specific goals for each classroom or course, establish student targets that operationalize each goal, and administer assessments to measure student attainment of the target. Aggregate student attainment of the target within a classroom can be attributed to the teacher and incorporated into evaluation systems. This process combines what Harris calls the "outcomes orientation of student test scores with the more subjective elements of classroom observations" (2012, p. 5).

The popularity of using SLOs in teacher evaluation systems can be attributed to federal policies such as Race to the Top (RTTT) and waivers to No Child Left Behind (NCLB) regulations, requiring that teacher evaluations systems include a measure of student growth for every teacher in order to evaluate their effectiveness (The Reform Support Network, 2014). These policies were influenced by research demonstrating that the quality of the classroom teacher is the most important school-based contributor to student performance (Chetty, Friedman & Rockoff, 2011; Kane & Staiger, 2008, Sanders and Rivers, 1996). combined with the observation that most teacher evaluation systems did not sufficiently differentiate teachers according to their ability to raise student performance (Weisberg et al., 2009; Vigdor, 2008; Aaronson et al., 2007; Clotfelter et al., 2007; Harris & Sass, 2008; Rivkin et al., 2005; Rockoff, 2004).

When statewide standardized assessments are available, value-added modeling (VAM), which refers to a class of regression-based models used to calculate a teacher's contribution to student learning based on standardized test score gains, is the primary method for holding teachers accountable. However, most state assessments are not administered in every grade and subject; in fact, nearly 70% of teachers teach in what are considered to be a non-tested courses (Prince et al., 2009). SLOs have emerged as a supplement to VAM-based metrics in these non-tested courses precisely because do not require standardized assessments in order to determine a teacher's contribution to student growth.

As I describe in greater detail below, SLOs are different from VAM metrics for several reasons. First, in order to align test and classroom content and given the lack of standardized components, SLOs constitute a process involving more steps than VAM metrics (Lacireno-Paquet et al., 2014; Marion et al., 2012). Second, they are intended not only to hold teachers accountable for student performance, but also to reinforce good teaching practices, such as goal setting and data-driven instruction (Slotnick et al., 2004). Third, SLOs can and do vary widely in terms of the level of standardization across states; therefore, they can look quite different from one state to the next (Lachlan-Hache, et al., 2013). And finally, teacher-level SLO scores do not explicitly control for outside factors which may impact student growth, such as student composition or school-level characteristics.

Despite differences in implementation and analytics, teacher-level SLO scores should validly and reliably distinguish teachers from one another in a teacher accountability framework, similar to VAM scores. In fact, the use of any metric based on

student growth in a teacher evaluation system must be seriously examined, in order to ensure that teacher scores are not systematically biased (AERA, APA, & NCME, 2014). This is particularly true of SLOs, since they can count toward as much as 50% of teachers' evaluation ratings (Lacireno-Paquet, et al., 2014), yet nearly no evidence exists to substantiate student and teacher-level score inferences for high-stakes decisions (Harris, 2012; Tyler, 2011). In particular, the extent to which the quality of assessments used and the choice of student growth targets affects student and teacher score inferences remains unclear.

Determination of Student and Teacher SLO Scores

To measure student SLO performance for each classroom/course, educators first create goals for each course that reflect the relevant content standards and cover the key concepts students should master (Marion and Buckley, 2011). Educators also create or select the assessments capable of measuring student achievement of these goals; these assessments should provide valid indicators of student performance (Steele et al., 2011). Finally, educators must set student targets required for students to demonstrate goal attainment. Two aspects must be considered when determining the target; the first is the method used to set the amount of student growth required of students, and the second is the rigor of the standard set for adequate growth. The tailoring of the goals, assessments, and targets to the curriculum and student makeup of each classroom or course is intended to ensure that the SLO is connected to the learning experiences and the ability of each student, thereby contextualizing each student's SLO score (Marion et al., 2012).

The goals, assessments, and targets can be determined by the classroom teacher, teams of teachers within the same course, school leaders or district leaders. States differ

in terms of the amount of teacher involvement in creating the SLO components. This reflects the tension among states between, on the one hand, wanting greater teacher engagement in order to increase teacher buy-in of the system and maximize the instructional benefits SLOs and, on the other hand, the need to ensure that SLO results are comparable across sites and valid as an accountability measure (The Reform Support Network, 2014; Lachlan-Hache et al., 2013). Regardless of the level at which the SLO is created, there is generally considerable variability across classrooms and courses within a state in terms of the assessments administered and the growth targets used (Lachlan-Hache et al., 2013).

Once assessments have been administered and graded (typically by the teacher), teacher-level SLO scores can be calculated. Teacher scores are often defined as the percentage of each teachers' students who meet or exceeded their SLO target. In order to determine whether the teacher is effective or not based on student SLO performance, a teacher's SLO score is generally translated into a three or four-point rating. For example, to be considered proficient, a teacher might need to receive a rating of 3 or higher, requiring at least 65% of the teacher's students to meet their target.

Literature Review

Research on student and teacher SLO scores comes primarily from evaluations of SLOs in pay-for-performance systems in three school districts: (1) Denver's Pro-Comp system, implemented in the 1999-2000 school-year (SY); (2) Austin Independent School District's (AISD) Reach program, implemented in the 2007-08 SY; (3) And Charlotte Mecklenburg's (CMS) TIF-LEAP initiative, implemented in the 2008-09 SY. Each evaluation includes data from interviews and surveys with teachers who have an SLO in

their classroom, and provides suggestive evidence that the SLO process is having a positive influence on teacher instruction based on teacher reports (Schmitt, Lamb, Cornetto and Courtemanche, 2013; Lamb, Schmitt, & Cornetto, 2010; Slotnick et al., 2013; Slotnick et al., 2004).

Evaluations of AISD's Reach Program and CMS's TIF-LEAP program also include quantitative findings comparing aggregate student performance on course SLOs with aggregate student performance results from state or districts assessments. Findings on the extent to which student SLO scores converged with other metrics of student performance in these districts were inconclusive. Across all studies conducted on AISD's program, results varied considerably across year of implementation and grade/subject examined (Schmitt, 2014; Schmitt, Lamb, Cornetto and Courtemanche, 2014; Schmitt, et al., 2013; Schmitt, 2011; Cornetto, Schmitt, Malerba, & Herrera, 2010; Schmitt, Cornetto, Malerba, 2009). For example, in the most recent year of the evaluation (2012-13), Schmitt et al. (2014) found that the correlations between the percentage of students who met their SLO targets and school-wide value added results across schools ranged from -0.05 in reading, to 0.14 in Social Studies, to 0.71 in science.

With CMS's pay-for-performance system, researchers found a positive relationship between student attainment of the SLO and student results from standardized test scores; however the size and strength of this relationship varied across years of implementation and grades, and was generally quite small (Slotnick et al., 2013). In the 2009-10 school year, elementary school students who met their SLO target demonstrated .11 to .15 higher standard deviation units on the ELA and mathematics state test then students who did not meet their SLO target. However, no statistically significant

relationship between SLO attainment and standardized test results was found at the middle school level. In the 2010-11 school year, a statistically significant relationship between student SLO scores and student performance on standardized assessments was found only in select grades and subjects.

Neither evaluation addressed why the association between SLO scores and other measures based on standardized assessments varied across years and subjects. A potential reason may be differences in the quality of student-level data being produced by the system across subjects and years. Issues with student-level data can include the quality of scale score data, the occurrence of score manipulation, and inconsistency in the standard required of students through the targets. The ways in which these issues impact student-level scores, particularly when used for high stakes teacher evaluation scores, is unclear. Therefore, in Chapter 1 of this study, I examine three major areas of concern with student-level data from one state: Assessment Data Quality, Score Manipulation, and Target Comparability.

The mixed and limited quantitative evidence on SLO scores also makes clear that a more thorough investigation into the overall validity (i.e., the degree to which the metric is measuring what it claims to be measuring) and reliability (i.e., the degree to which the metric is consistently measuring what it claims to be measuring) of teacher SLO scores is needed given that these scores will be used to make high-stakes decisions regarding teachers (Harris, 2012; Tyler, 2011). Such evidence is required of metrics used to measure teacher performance in high-stakes settings (AERA, APA and NCME, 2014) and has been collected for value-added scores (see for example, Loeb and Candelaria, 2013; Hill et al, 2011; Bill and Melinda Gates Foundation, 2010). Moreover, research on

teacher SLO scores must also take into account how the features that make up the SLO score – the assessments and growth targets – affect the interpretability of teacher scores. In Chapter 2, I examine evidence regarding the validity, reliability and comparability of teacher SLO scores in one state.

Site Context

Data for both chapters of this study come from one state that implemented SLOs as part of their teacher evaluation system under an RTTT grant. The 2012-13 school year was considered a pilot implementation year for RTTT districts, whereby districts were responsible for developing an SLO for every non-tested course in which a teacher did not have a growth score based on standardized assessments. SLO results from this year were not used for high-stakes considerations. The 2013-14 school year was the first full year of implementation for these districts. It was initially considered a high-stakes year for RTTT districts only, such that 2013-14 SLO scores would be factored into the state's teacher evaluation scores in the following year and would, by law, be required to inform human resource decisions.[1] While federal guidance has since allowed states to delay high-stakes implementation, teacher evaluation scores in this state will continue to be used as the basis for merit-pay bonuses, and importantly, at the time of implementation, districts were under the impression that the scores would be used for high stakes decisions.[2] In neither year were SLOs considered high-stakes for students; however in most districts,

[1]Final calculations of teacher SLO scores in this state cannot be completed prior to the end of each school year, given when the post-test is administered. Thus, the SLO score included in each teacher's final evaluation score comes from the prior school year (e.g., teacher SLO scores from 2013-14 will be included in 2014-15 teacher evaluations).

[2] Some principals have indicated they will use the 2014-15 final evaluation scores, which include 2013-14 SLO results, for teacher placement decisions and to inform human resource decisions; however this is not mandated by the state.

teachers are allowed the option to include the post-score assessment results in final course grades, and principals may use the results for placement in gifted or remedial courses in the following year.

A teacher's SLO score is calculated as the percentage of all students taught who met their SLO target. To set standards for teachers that factor into an evaluation system, this percentage is transformed into a four point rating (1-4), using pre-established cut-points.[3] For the majority of teachers with an SLO score, this metric is the only measure of student growth included in their evaluation, and counts for 50% of their final effectiveness rating. The other 50% is based on classroom observations conducted by school leaders based on the Classroom Assessment Scoring System (CLASS) protocol, developed by Pianta and colleagues (Pianta, Hamre, Haynes, Mintz, &La Paro, 2006).[4]

The state has mandated the centralized creation of SLOs at the district level, including the creation of the goals, the selection/creation of the assessments and the selection of the student growth target. Therefore, every teacher within a district/course administers the same assessment and applies the same student growth target formulation to student scores to determine whether each student met the SLO. Although teachers do not create an individual SLO for their classroom, teams of teachers typically work with district-level personnel to develop the SLO for a particular course.

The assessments used for SLOs in this state consist of both commercially-developed and locally-developed assessments, including tests created using an item bank

---

[3] At the time of this study, this state had not yet determined the final thresholds for determining teacher effectiveness.

[4] CLASS is a classroom observation protocol developed to assess teacher ability around the domains of instructional support, emotional support and classroom organization. CLASS measures 10 specific standards: professional knowledge, instructional planning, instructional strategies, differentiated instruction, assessment strategies, assessment uses, positive learning environment, academically challenging environment, professionalism, and communication.

established by the state. Local assessments are the norm; however little is known about the quality of these assessments since there are no technical reports detailing test reliability. Commercial assessments, such as NWEA and STAR assessments, are used typically in early-elementary ELA and Mathematics courses in this state. Every SLO includes a pretest and a posttest in order to measure student growth, and often times, the pre-assessment is the same or a portion of the post-assessment. Assessments include multiple-choice items as well as open-ended items.

In addition to selecting the assessment for each course, districts are expected to choose from one of four different types of growth target formulations, each of which set target scores as a function of student prescores. Districts can adjust the rigor of the target formulation as they see fit. Target scores for every student within a course must be calculated in the same manner, therefore every student in a district/course with the same prescore should have the same target score, regardless of other observable characteristics (i.e., whether the student is classified as having a disability).

The four target types districts can choose from include: (1) an individual growth target; (2) a tiered growth target; (3) a uniform growth target; and (4) a categorical growth target. The individual target requires students to reduce the gap between their prescore and the maximum number of points on the test by a certain percentage. For example, if students in one course are required to grow by 25% of their potential growth (the difference between their prescore and the maximum score on the test), on a 100 point test, a student who scores a 50 on the pretest will be required to get a 62.5 on the posttest in order to meet his/her target score. A tiered target sets cut-points on the pretest and requires students scoring in each tier to improve by a set number of points on the posttest,

with the number of required points decreasing for each successive tier. With a uniform target, every student must increase from their prescore to their postscore by an equal number of points. This is often given with rubric-based assessments that are typically designed around 4 to 6 score points, whereby students are expected to gain 1-2 points from pre- to posttest. The final target type is a categorical growth target, whereby cut-scores are imposed on the pretest and posttest based on meaningful categories of scores, and students falling in one category on the pretest are expected to reach the next successive category on the posttest.

Because teachers often teach multiple courses, approximately 20% of teachers with an SLO also teach a course with a state test and will therefore have a teacher effectiveness measure known as a mean Student Growth Percentile (MGP) score. Student Growth Percentiles (SGPs) are similar to VAM metrics in that each student is compared to others with a similar score trajectory. SGPs differ from VAM metrics in that the SGP metric is based on a non-parametric form of regression, quantile regression, in order to derive a student percentile ranking for each student based on the distribution of their current scores relative to other students with similar test scores (Betebenner, 2009).[5] Further, this method is geared toward describing student growth rather than teacher effectiveness; as such, the mean SGP among all students taught by a given teacher is typically used as the summary metric at the teacher level.[6] For teachers with both an

[5]For a more technical description of the differences between VAM and MGP metrics, see Castellano and Ho, 2013)

[6] Teacher scores from mean SGPs have been found to be highly related to scores based on different specifications of value-added models (Goldhaber and Theobald, 2013); however, given that they do not control for student background characteristics, they tend to slightly favor teachers in more advantaged classrooms compared to VAM models which due control for background characteristics.

SLO and MGP score, the "student growth" component of their evaluation will be based equally on both scores.

In order to dig deeper into the nuances of the assessments and growth targets and provide greater context surrounding the implementation of SLOs, the dataset is limited to a sample of districts and courses, which were chosen to maximize the number of students and teachers with SLO scores. Districts in the sample are not identified by name in order to maintain confidentiality; instead they are referred to as Districts A-H. These districts span the need/local/performance spectrum (see Table 1). Each chosen district is an RTTT district, and therefore contains SLO data from 2012-13, a non-high-stakes year, and from 2013-14, an intended high-stakes year. The courses are limited to core academic courses that are similar across districts, including: third grade English Language Arts (ELA) and Mathematics, and 10th grade Literature/Composition, World Literature, British Literature, Algebra, Chemistry, American Government, and World History.

[insert Table 1]

In Chapters 1 and 2, I present findings on student-level analyses and teacher-level analyses, respectively. Both chapters provide the methods, results and discussion on each respective set of research questions. I conclude this dissertation by discussing which factors states and districts should focus on in order to improve the validity of student and teacher SLO scores.

Tables

Table 1. Demographic and performance data, for each sample district and the state.

District	Locale	# schools	# students	# teachers	%SWD	%ELL	% nonwhite	% prof math
District A	City, midsize	59	35,842	2,603	10.2%	1.7%	57%	79.4%
District B	Suburban, large	41	38,774	2,377	10.6%	3.9%	8%	93.0%
District C	Suburban, large	66	51,018	3,242	9.4%	7.1%	73%	79.1%
District D	Suburban, large	145	98,088	6,650	8.5%	9.9%	79%	75.8%
District E	Suburban, large	134	162,370	10,323	11.1%	11.7%	31%	91.5%
District F	Rural, fringe	34	26,261	1,662	10.3%	14.8%	20%	85.2%
District G	Rural, fringe	51	39,909	2,628	13.2%	1.7%	33%	87.6%
District H	City, midsize	61	32,231	2,308	11.3%	1.5%	60%	76.7%
State	-	2,246	1,639,077	110,429	10.4%	5.1%	55.9%	84.7%

Notes. District demographic data come from the NCES website, based on the 2011-12 school year. State demographic data come from a document produced by the state department of education, based on the 2011-12 school year as well. District and state student performance data come from the state AYP website, based on the 2010-11 school year.

Chapter 1: An Evaluation of the Interpretability of Student Learning Objective Results Based on Student Assessments and Growth Targets

Introduction

In Chapter 1, I investigate three factors which can affect the interpretability of student-level SLO scores. The first is the quality of assessment data, based on evidence from student prescores and postscores; specifically I look for evidence of anomalous distributions and instances where courses were student prescores fail to predict student postscores. Second, I look for a very specific type manipulation of student postscores among students whose score falls right below their target score. And third, I examine the sensitivity of student scores to variation in student growth target formulations and rigor. I discuss each issue in greater detail below.

Assessment data quality

No research to date has examined whether the assessments used for SLOs have sufficient technical quality to measure student achievement (Gitomer and Bell, in press; Herman, Heritage and Goldschmidt, 2011; Steele et al, 2011). While standardized assessments typically contain technical reports which provide evidence on the validity and reliability of the assessment, most of the assessments being created for SLOs are newly created, or at the very least, newly implemented, and are not accompanied by reliability and validity analyses. Based on a survey of states implementing SLOs, the assessments used vary widely, from commercially-created tests, such as state standardized assessments and vendor-developed assessments, to locally-developed tests, such as district end-of-course-assessments, as well as school or classroom-developed

measures (Lacireno-Paquet, et al., 2014; American Institutes for Research, 2013; Buckley and Marion, 2011).

In general, student scores from standardized assessments follow a normal distribution, without extreme skew in either tail of the distribution (Koretz, 2008). Bi- or multi-modalities can indicate inconsistencies with how the assessments are administered or graded, and considerable skew (greater than -1/+1) can indicate a mismatch between the ability of students taking the test and the rigor of the test (Ho and Yu, 2014). If anomalous or skewed test-score distributions occur, student scores may fail to provide an accurate or fair representation of student ability in the given content domain.

Investigating assessment quality is further complicated by the use of a pretest-posttest format, which has been introduced in most sites using SLOs due to federal policies requiring the measurement of student growth over time (Lacireno-Paquet, et al., 2014; The Reform Support Network, 2014; Goe and Holdheide, 2011; USED, 2011). If different assessments are used and the items on each have not been placed on the same scale, i.e., test scale linking – a statistical process seldom undertaken by districts, it can lead to over- or underestimation of student growth when growth is calculated as a gain score (Castellano and Ho, 2013; Marion et al., 2012; Diaz-Bilello, 2011; May et al., 2009). An investigation of the correlation between student prescores and postscores can indicate the extent to which the pre and post-tests are aligned: with standardized assessments, these correlations are generally quite strong (0.6-0.8) (Cole et al, 2011). Low correlations, on the other hand, typically suggest lack of variability in one or both of the tests due to floor effects (i.e., most students answered the questions incorrectly) or ceiling effects (i.e., most students answered the questions on the test correctly), that the

two tests are not measuring the same construct, or that measurement error in one or both of the tests is attenuating the reliability of student scores (Castellano and Ho, 2013; Koretz, 2008; Cole et al., 2011).

Score manipulation

Another issue related to the assessments that can affect the quality of student SLO scores is that of score distortion, which follows from Campbell's law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1975, p. 35). Score distortion refers to instances where test scores do not accurately represent student performance due to specific actions undertaken by educators to improve scores, such as a narrowing of the curriculum or explicit test preparation, each of which are generally accepted– although not desired – actions in an age of high stakes testing (Koretz, 2008; Koretz & Hamilton, 2006; Koretz & Barron, 1998).

Researchers and school administrators have found evidence of more overt score manipulation, namely, cheating, whereby teachers have directly altered student scores on the assessment (Jacob and Levitt, 2003; Dee, Jacob and McCrary, 2011; Will, 2014). Dee and coauthors examined score manipulation on the New York Regents Exam, where teachers are responsible for grading the student assessments that are used to determine high school graduation. Based on a comparison of the distribution of scores right below and above each performance threshold score required for graduation, the authors concluded that school-based grading of the exams incentivized officials to artificially

increase scores for 3% to 5% of students who were on the cusp of meeting the required score.

Score manipulation is a major concern with SLOs since, like with the NY Regents exam, teachers generally administer and grade student pre- and posttests (Marion et al, 2012). Given that results from SLO assessments will directly factor into teacher evaluation scores, this produces a strong incentive for teachers to alter scores in order to protect their jobs (Steele et al, 2011; Koretz, 2008). Moreover, gain scores, which are typically calculated with SLOs, are particularly susceptible to score manipulation, since the depression of prescores *and* inflation of postscores can inflate score results (Castellano and Ho, 2013).

Choice and calculation of student growth target

In addition to the quality of assessment data and the scoring of the assessments, the choice and calculation of student growth targets can impact the interpretation of student SLO performance. With SLOs, student growth target formulations may vary across students, classroom and districts which can limit the comparability of student scores. The comparability of student scores under different growth targets is conceptually similar to a line of research carried out on accountability systems propelled by NCLB, where researchers found that the interplay between specific growth models and state contexts altered inferences about student growth and substantially impacted the classification of schools meeting their growth targets, even when holding the underlying data constant (Goldschmidt, Choi & Beaudoin, 2012; Hoffer et al., 2012). Yet, no research has examined how variation in SLO growth model formulations affect inferences about student SLO performance.

A secondary concern with growth targets can arise since the calculation of the SLO target score is not necessarily automated. In fact, typically, it is the classroom teacher that calculates each target score based on student baseline performance (i.e., the pretest scores) (Lachlan-Hache et al, 2012). As such, the extent to which errors in growth target calculations occur and impact student SLO attainment is unknown.

Research Questions

There are many issues concerning student SLO scores that cannot be addressed –or addressed fully - due to data limitations (i.e., I do not have item-level data to examine the internal consistency of student assessment scores, nor do I have the actual test items to evaluate the alignment of the rigor and content of pre and post-assessments). Instead, using student test score data from one state, wherein assessments and growth targets are chosen by the district and applied uniformly within a district/course, I address the three research questions presented below. When appropriate, I compare findings on each research question from a non-high-stakes year to an intended high-stakes year to determine if any issues that arise are corrected prior to student scores being used presumably for high-stakes decisions.

1. *Based on student prescore and postscore data, to what extent:*
    a. *Do the assessments contain anomalous test score distributions, including bi-modalities, scores beyond the maximum number of points on the test, extreme skew or extreme spikes in the tails of the distribution?*
    b. *Are student prescores predictive of student postscores?*
2. *Is there evidence of manipulation of student postscores near the target score in a high-stakes year?*
3. *How does the student growth target affect the classification of student SLO attainment due to:*
    a. *Differences in the type and rigor of growth target formulations used across courses and districts?*
    b. *Potential errors in target score calculations?*

Data and Method

Data

A state-produced dataset containing student-level SLO data from the 2012-13 and 2013-14 school year was used to address each research question. The data include student prescores, postscore and target scores. Using this data, I created an indicator variable for whether each student met his/her growth target (*SLO_met*), based on whether each student's postscore was equal to/greater than his/her target score. Teacher, course and district IDs were used to classify the teacher, course and district for each student.

Sample

The sample includes a total of 117,834 students with SLO scores in the 2012-13 school year, and 157,600 students with SLO scores in 2013-14 school year. Note that a student may be counted more than once if he/she has an SLO score from more than one course. Table 2 displays the frequencies of students in each district/course in each year. In general, the number of students with an SLO score from each district increased from 2012-13 to 2013-14 as districts moved towards full implementation in the latter year.[7]

[insert Table 2]

Data analysis

Assessment data quality. Data visualization and descriptive statistics are essential first steps when analyzing secondary data based on assessment results, and have been used extensively to examine results from state standardized assessments (Ho and Yu, 2014; Cole et al., 2011). Therefore, to deduce potential issues with assessment quality, I first looked for clear evidence of bi- or multi-modal score distributions, as well as student

[7] The primary exception to this is District F, which only submitted to the state student scores for a maximum of two courses for each teacher in 2013-14.

scores beyond the maximum number of points on the test, both of which can suggest inconsistencies with test administration or grading. I also examined the data for spikes in the tails of the distribution and calculated skew statistics, in order to determine whether there was mismatch between the difficulty of the test and the ability of students taking the test. Finally, I calculated Pearson correlations of student prescores and postscores in order to determine if student prescores were predictive of student postscores, an important factor when growth scores are being used to determine student performance. The results were compared to those typically found with standardized assessments in order to judge the overall quality of the assessments used for SLOs.

Score manipulation. To examine evidence of score manipulation, I used the McCrary density test, developed to investigate manipulation of the running variable in a regression discontinuity framework (McCrary, 2008). It is based on the supposition that absent any manipulation, the density of the running variable is continuous. As such, it assumes unidirectional manipulation (i.e., manipulation of scores in one direction), which is what we'd expect to find if teachers are engaging in score distortion by increasing student scores above the target score.

To determine if there is is a significant difference in the density of scores across the point at which manipulation is expected, the McCrary density test uses a local linear density estimator (Cheng et al, 1993; Cheng, 1994). Student scores are first binned into non-discrete cells (i.e., bins that span at least two values of the variable of interest): the height of the histogram bins (based on the midpoint value) is used as the independent variable and the number of observations within each bin is used as the outcome variable.[8]

[8] Bandwidths and bin-sizes are estimated automatically by the program (see McCrary, 2008); however I adjust them in additional sensitivity analyses.

Two separate regressions are conducted, one on either side of the potential discontinuity, using local linear smoothing, with the most weight given to histogram bins nearest the potential discontinuity. The log difference of the intercepts of each regression is the estimate of interest.[9]

Since a student's target score, including the point of discontinuity, will vary for each student based on his/her prescore, I created a running variable *post_target*, which is equal to each student's postscore minus target score. When *post_target* is equal to 0, it indicates that a student's postscore exactly equals his or her target score. Thus, the frequencies of *post_target* I am most concerned with are those that occur at zero (i.e., the breakpoint) and to the immediate right and those that occur just to the left of zero; in other words, the frequencies of students who just meet their target score compared to students who just fail to meet their target score. [10]

I make use of the difference in stakes in each year by comparing the local linear density estimate in a high-stakes year with the local linear estimate in a low-stakes year.[11] To do so, I first calculated estimates by year and district/course, for all students who are taught by a teacher included in the sample. I then calculated a z-ratio by taking the difference in estimates from both years and dividing it by the standard error of the difference.[12] If the z-ratio>1.96, it suggests a significant difference in the frequencies of

---

[9] The derivation of the standard errors of the estimate of interest can be found here: http://eml.berkeley.edu/~jmccrary/DCdensity/DCdensity.pdf

[10] For the McCrary density test to run, there needs to be a sufficient range of scores on either side of the discontinuity. For example, if within a district/course, all students meet or exceed their target, a local linear estimate on the frequency of scores to the left of the breaking point cannot be estimated.

[11] This analysis is akin to a difference-in-difference test, where the first difference is applied to scores to the left and right of the breaking point, and the second difference is applied to log estimates from a high stakes and low stakes year.

[12] The standard deviation of the difference=$\sqrt{se_{year1}{}^2 + se_{year2}{}^2}$

students across the breakpoint between 2012-13 and 2013-14, based on a two-tailed test with an alpha-level of .05 (i.e., 95% confidence level).

Choice of student growth target. To examine how the type and rigor of the target impacts inferences of student performance, my analyses closely follow Hoffer et al. (2011), in which alternative growth target formulations are applied to the same underlying data to see how student SLO attainment changes. Two sets of alternative targets were applied to the data. The first set is based on each of the four target formulations used by districts, including an individual growth target ($Target_{Indiv}$), a tiered growth target ($Target_{Tier}$), a uniform growth target ($Target_{Uni}$), and a categorical growth target ($Target_{Cat}$). To hold the rigor of the target constant, each target formulation was constructed so that the same percentage of students in the 2012-13 sample met the target under each alternative target formulation (see Table 3).[13]

[insert Table 3]

Next, three alternative targets were created, whereby the amount of growth expected of students from their prescore to the maximum number of points on the test (i.e. potential growth) varied, but the type of growth target was kept constant (i.e., an individual growth target).[14] $Target_{Indiv_{25}}$ requires students improve their postscores by 25% of their potential growth, $Target_{Indiv_{50}}$ requires students to improve their postscores by 50% of their potential growth, and $Target_{Indiv_{60}}$ requires students to

[13] On average, 52% of students in 2012-13 had postscores that met each alternative target score. Note this percentage was based on applying each target formulation to every student in the sample of district/courses.
[14] I chose an individual target for this simulation since it allows for the most straightforward conception of rigor, shifting the percentage growth required. Unlike with an indidvual target, rigor for tiered and categorical targets can vary in several ways, including the amount of growth required of each tier/category, the difference in the amount of growth required from one tier/category to the next, the range of scores that make up each tier/category, and the number of tiers/categories included.

improve their postscores by 60% of their potential growth. These variations were commonly used by districts in year one and year two, and represent a shift in the rigor of the target, since lower percentages of potential growth requires less growth among students than higher percentages of potential growth.

A total of seven alternative growth target formulations were created, four that varied in type but contained similar average rigor, and three based on the same target type but with escalating rigor. I applied each alternative target to the underlying data in the sample and created seven new indicator variables for each student, indicating whether each student met his/her SLO under each alternative target type. Therefore, for every student, the dataset contained eight indicators of student SLO attainment: one based on the target set by the district ($Target_{District}$), one for each alternative target formulation whereby target type varied but target rigor was held constant ($Target_{Indiv}, Target_{Tier}$, $Target_{Uni}, Target_{Cat}$), and one for each alternative target formulation whereby the target type remained the same but rigor varied ($Target_{Indiv_{25}}, Target_{Indiv_{50}}$, $Target_{Indiv_{60}}$).

A comparison of performance across the first set of alternative targets ($Target_{Indiv}, Target_{Tier}, Target_{Uni}, Target_{Cat}$) indicates how student performance is affected by the growth target while holding the rigor of the growth target constant. A comparison of student performance across the second set of alternative growth targets ($Target_{Indiv_{25}}, Target_{Indiv_{50}}, Target_{Indiv_{60}}$) indicates how student performance is affected by the rigor of the growth target while holding the target type constant.

Calculation of student growth target. Recall that within a district/course, all students were given the same growth target formation; therefore all students in the same

course with the same prescore should have the same target score. However, given that target scores are typically calculated by teachers, rather than automated, target errors can and do occur. Two-way scatter plots of student target scores by prescores within a district/course reveal those students with the same prescore but different target scores.

What is not entirely intuitive is which students have correct target scores and which students have incorrect target scores, since the actual target calculation used by districts is unknown. With an individual or uniform target, where the formula is the same across students, one can deduce the target formulation being applied to the majority of students and determine which students' target calculations were inconsistent with this formulation. With a tiered or categorical target, the amount of growth required is different (and unknowable by the researcher) for each successive category/tier. I therefore classified the correct amount of growth for each tier/category as that which occurred more frequently among students in the same district/course with the same prescore. For example, among students in District G's third grade mathematics courses who received a 4 on the pretest, 64% were given a target score of 5, while 36% were given a target score of 6; I therefore classified a target score of 5 as the correct target score. Once incorrect target scores were identified and corrected based on a new variable, $Target_{correct}$, I calculated the number of students who would switch SLO attainment classification based on whether their postscore was equal to or greater than their corrected target score.

Findings

Assessment data quality

I compared the quality of assessment data found in 54 district/course combinations in the 2012-13 year to the quality of assessment data found in the 59 district/course combinations in the 2013-14 school year (recall that a greater number of

courses contained SLOs in 2013-14 compared to 2012-13). District/course combinations in both years were limited to those in which at least 200 students took the assessment, given the issues that can arise with small sample sizes.[15]

Based on informal conversations with district and state personnel, every district in the sample made some sort of modification to at least some of their assessments between 2012-13 and 2013-14. District A, for example, modified their assessments slightly by removing some items due to item difficulty or reliability, while District D and District B introduced entirely new assessments in third grade Mathematics and ELA.[16] District E created and administered SLOs for six additional subjects in the sample, and modified existing assessments in order to align with the Common Core of State Standards (CCSS).

Issues with test administration and scoring. I first looked for bi- or multi-modal distributions and student scores beyond the maximum number of points on the assessment (i.e., illegal scores) in year one and year two, which can suggest issues with test administration and scoring. Issues with test administration/scoring can place student scores within a district/course on different score-scales, thereby making any inferences of attainment of these students invalid.

Bi-modal score distributions. In total, twelve assessments in the sample of district/courses in 2012-13 had scale score distributions that were clearly bi-modal,

---

[15] For example, in 2013-14, District F only reported scores for teachers in a maximum of two courses, which reduced the number of students included in the sample for District F in year two compared to year one. Requiring a student n-size of at least 200 students ensures that the student scores are representative of the entire sample of students taking the test.

[16] District D switched their assessments from 100 point locally-developed assessments to 1400 point commercially-developed Renaissance Learning (STAR) assessments, and District B switched their third grade Mathematics and ELA assessments from commercially-developed assessments to district-developed assessments. While District D made no changes to assessments in the remaining subjects, District B made significant modifications to the assessments used at the high school level, replacing public domain items with locally developed items.

suggesting inconsistent test administration or inconsistent test grading.[17] An example of inconsistent test administration procedures occurred with District D's British Literature post-assessment. This assessment contained a multiple choice section with twenty questions worth two points each for a total of forty points, and an essay section evaluated on a three to twelve-point scale, with each point worth five points towards the student's total score. A histogram of student scores from this assessment is displayed in Figure 1, Panel A, which shows clear evidence of a bi-modal distribution, with the first mode ranging from two to forty points, and the second mode ranging from 41 to 60. Given the scoring framework, the distribution of postscores should contain both even and odd scores. However, scores in the first modality are primarily even numbers, suggesting that these scores come from classrooms in which only the multiple choice section was administered. In fact, every student in five out of the twenty schools administering this assessment, comprising 27% of all student scores on this assessment, had an even test score that did not exceed a score of forty. This suggests that in these schools, teachers did not administer the essay portion of the assessment.

An example of inconsistent test scoring procedures occurred with District F's British Literature assessment. This assessment was a 100-point, locally-developed assessment. A histogram of student postscores from District F's British Literature assessment is displayed in Figure 1, Panel B. A bi-model distribution is evident, with one mode ranging from zero to four points, and the second mode ranging from ten to 98 points. Greater investigation into the first mode reveals that these scores came from just

[17] In all but two cases, bi-modal distributions were found with both the pre and post-assessment within a given district/course, suggesting that any inconsistency with test administration or scoring that occurred with the pretest also occurred with the posttest.

two teachers in one school, whereby every student received a score of between zero and four points, on the posttest as well as the pretest. These scores constitute 13% of all postscores and suggest that these two teachers graded the test on a four-point scale instead of a 100-point scale.[18] A similar investigation of scores in the remaining five district/courses that displayed evidence of bi-modal distributions suggests the occurrence of inconsistent test administration/scoring as well.

[insert Figure 1]

In the 2013-14 school year, only six assessments exhibited a bi-modal scale score distribution: District D's third grade ELA and British Literature pretest and posttest and District G's Chemistry and Algebra pretest. The reduction in bi-modal test score distributions from year one to year two of implementation suggests greater consistency in how teachers were administering and grading the assessments over time.

Scores that occur outside of the boundary of maximum points on the test. A second indication of administration/scoring errors occurs when student test scores exist outside of the boundary of the maximum number of points on the test, which I refer to as illegal scores. If the illegal score occurs on the posttest, the student will be guaranteed to meet his/her target score regardless of actual ability since no target score is greater than the maximum number of points on the test.

In 2012-13, illegal test scores occurred with a total of eight assessments in the sample: four pre-assessments and four post-assessments. However, the number of students with illegal scores was quite small: one to two cases from each assessment. For example, on District A's third grade mathematics exam, a commercially-developed exam

[18] I find a similar occurrence with one of these teachers on the World Literature assessment in District F.

by NWEA that had score range of 120 to 300 points, one student had a score of 493, far outside the total number of points on the test.

In 2013-14, illegal test scores continued to occur in eight assessments. As in 2012-13, the number of illegal scores was quite small (1 to 2 students).[19]

Mismatch between assessment difficulty and student ability. Distributions with extreme skew and large spikes in the frequency of student scores at either tail of the score distribution can indicate mismatch between assessment difficulty and student ability Mismatch can produce floor and ceiling effects, defined by Ho and Yu as "insufficient measurement precision to support desired distinctions between examinees at the upper and lower regions of the score scale, respectively" (2014; p. 10), and ultimately lower the reliability of student gain scores. In a teacher evaluation system, this issue can create bias in teacher scores if, for example, some teachers' students cannot show as much growth as others due to the difficulty of the assessment (Koedel and Betts, 2010; Diaz-Bilello, 2011).

Extreme skew. Skew statistics measure the asymmetry in the distribution based on the distance of each observation from the mean.[20] Assessments with skew statistics of 0 are considered perfectly symmetrical; most standardized assessments contain skew statistics between -1 and +1 (Ho and Yu, 2014). Extreme negative or positive skew on an assessment *can* indicate that the SLO assessments are not distinguishing sufficiently

---

[19] In several courses in District E, illegal scores appeared to be a much larger and more systematic issue; subsequent conversations with district and state personnel suggest that it was likely the result of problems when the data was transferred from the district to the state. Given that this was not a function of the administration/scoring of the assessment or quality of the assessment itself, I did not include these assessments in my count of assessments with data quality issues, and removed the illegal scores from subsequent analyses. I also checked state data against district data and when there were discrepancies, used district data instead.

[20] The formula for a skew statistic is: $s = \frac{\sqrt{n(n-1)}}{n-2} \frac{\frac{1}{n}\sum_i (x_i - \bar{x})^3}{(\frac{1}{n}\sum_i (x_i - \bar{x})^2)^{3/2}}$.

among higher scoring or lower scoring students, respectively. Because assessments in the sample are generally locally-developed, I use a lower standard to classify extreme skew: <-2 and >+2.

Important to note is that while skew statistics can indicate floor or ceiling effects, they are not necessarily indicative of problems with student test scores (Ho and Yu, 2014). For example, skew statistics can be large if there are a small number of extreme outlying values in the data, but these may nonetheless be valid observations from the population. For this reason, I visually examine scale score distributions in addition to calculating skew statistics in order to classify mismatch.

Figure 2 contains a histogram of district/course skew statistics for test scores from each of the 54 pretests (panel A) and posttests (panel B) in the sample in 2012-13. Eight assessments display extreme positive skew: seven pre-assessments and one post-assessment. Extreme positive skew on the pre-assessment is likely attributable to the fact that districts typically use the same end-of-course assessment (or a portion thereof) for the pretest as they do for the posttest: an assessment developed as an end-of-course assessment will likely be too difficult for students at the beginning of the course. Extreme positive skew on the post-assessment was found in District A's third grade Mathematics course; however, this is an example of a misleading skew statistic since the statistic was unduly influenced by an illegal score of 493, discussed previously, outside of the range of possible points. No assessment in 2012-13 demonstrated extreme negative skew.

[insert Figure 2]

An example of extreme positive skew occurred with District H's World History pre-assessment, which has a test score scale of 100 points. As displayed in Figure 3, the shape of the distribution of prescores from the assessment is highly asymmetrical (s=2.7), with nearly 30% of students scoring at the lowest level (1) on the exam, every student scoring less than 70, and a mean score of seven. In this case, the shape of the distribution indicates that as a pre-test, the assessment was too difficult for the group of students it was administered to.

[insert Figure 3]

Only two pre-assessments displayed extreme positive skew in 2013-14: District A's Algebra pretest and District G's American Government pretest.[21] The reduction in number of tests with extreme skew from 2012-13 to 2013-14 suggests an improvement in the match between assessment difficulty and the ability of students from year one to year two of implementation.

Spikes in the tail of distribution. Typically, large frequencies of students falling at either end of the scale distribution suggests that the test is either too hard or too easy, given the ability of students it is administered to, which can lead to floor or ceiling effects. I classify spikes as those in which 10% of more of students have the same discrete scale score at either tail of the score distribution. This is an arbitrary classification; however, given that we would expect to see a normal distribution of student scores with a small percentage of scores falling at the lowest and highest value on

[21] Two additional assessments contained extreme positive skew; however, in the first case, District F's third grade Mathematics pretest, it was due to a few illegal scores; and in the second case, District G's third grade mathematics posttest, it was due to the fact that the test contained a total of eight score points, and most students scored a 1 or 2.

the assessment, I would argue that 10% or more of student scores falling at a single score point at the tail of the distribution is a clear indication of test mismatch.

In 2012-13, a total of 26 assessments, eighteen pre-assessments and eight post-assessments, contained a large spike in the frequency of scores at the left tail of the test score distribution. For example, in several courses in District C, large frequencies of students scored the lowest possible number of points on the pretest *and* the posttest (10% to 60% of students received one out of 100 points on each pre- and post-assessment in each course). In every subject in District D, 30-60% of students received a zero on the pre-assessment. Likewise, in District F's third grade Mathematics pre-assessment, which was based on a 5-point scale, 98.5% of students who took the assessment scored a one, the lowest score.

In none of the 26 assessments does this large spike in scores appear to be the result of data-entry issues. Students who scored the lowest number of points on the pre-assessment were all given the lowest possible target score as well; if a range of target scores was found, it would suggest that the prescore was a placeholder for missing data. Interestingly, the large spike in the frequency of student scores at the low end of the test score distribution does not consistently result in an extreme positive skew statistic. Therefore, had I relied solely on skew statistics to identify mismatch, I would have missed several occurrences. This illustrates the need to visually examine scale score distributions in addition to relying on skew statistics when determining mismatch.

In year two, there continues to be a large number of assessments (n=14) with students scoring at the tail of the distribution. However, this was primarily due to District D, where every high school course contained a pre and post-assessment with a large

percentage of students scoring a zero. [22] Large spikes in scores also occurred at the maximum number of points on District A's American Government and World History post-assessment.

The relationship between student prescores and postscores. In order to investigate whether student prescores predict student postscores, I calculated a prescore-postscore correlation for each district/course in the sample. Literature on standardized assessments typically finds correlations of student prescores and postscores of 0.6 to 0.8, suggesting the assessments are aligned and reliable (Cole et al., 2011; May et al, 2009). With standardized assessments, postscores are typically based on the current year's spring test, and prescores are generally based on the prior year's spring test.

With SLO assessments, it may be reasonable to expect correlations among pre- and post-assessments at the upper end of those found with standardized assessments, since, unlike with standardized assessments, the same or similar items are typically used for each pre and post-assessment in this state, and test administration is fall-to-spring, covering one school year, rather than a spring-to-spring administration. A low pre-postscore correlation would suggest that student prescores are not predictive of their postscores which may be attributed to unreliable tests, floor/ceiling effects that limit the

[22] Spikes in the test score distribution occurred with two additional pre and post-assessments in District D: third grade ELA and third grade mathematics. However, I did not include these in my count of assessments with serious data issues, since, as suggested by state and district personnel, this was likely an issue with uploading student scores into the state system. This was evident by the fact that among students who had a prescore of zero, there was a range of target scores (suggesting target scores were based off of unobserved prescores that varied). I also removed these cases prior to calculating skew statistics and pre-postscore correlations, since they did not appear to represent students' true prescore. Additionally, in District A's British Literature course, 20% of students scored a 20 on the pre-test and 8% scored a 20 on the posttest. However, district and state personnel indicated that this was likely due to the format of the assessment and not due to quality issues or grading issues with the assessment.

range of scores, or a lack of comparable constructs. I classify extremely low correlations as those less than 0.2 (including negative correlations).

Figure 4 presents a histogram of pre-postscore correlation coefficients based on the pre and post-assessments administered in 2012-13. The pre-post correlations range from -0.2 to 0.88. Eleven district/course combinations in the sample had low pre-post correlations (less than 0.2) or negative correlations. Assessments in nearly all of these courses exhibited additional quality issues, suggesting that the low pre-post correlation may also be driven by assessments that, in general, have poor quality.

[insert figure 4]

There does not appear to be an improvement in the alignment of pre and post-assessments from year one to year two; a total of fifteen courses contained low pre-post correlations in 2013-14. The fact that more courses in year two contained low pre-postscore correlations than in year one could be due to the introduction of new SLO courses in 2013-14 and/or changes to the assessments from year one to year two that did not improve the reliability of the assessments. On the other hand, low correlations in year two could be the result of issues such as teaching to the test, or re-testing incorrect items throughout the year, actions that would serve to reduce the predictiveness of student prescores. One piece of evidence to support the latter hypothesis are the high gain scores found in courses with low pre-post correlations in year two. In fact, of the 15 courses with low pre-post correlations, all but those from District D had above-average standardized gain scores (based on a sample average of 1.7).

Comparison of assessment data quality in year one and year two. I classified district/course combinations as containing an assessment with data quality issues if the

assessment contained at least one of the following: (1) clearly bi-modal score distributions, (2) skewed distributions (<2/<-2) or distributions containing large spikes in the tails (10% of students or more), and (3) small or negative pre-postscore correlations (<.2). Illegal scores were not included in the count of assessments with serious issues, given that the number of illegal scores found was generally quite small (1-2 students) on any given assessment and would therefore have a minimal effect on a teacher's overall score.

An important question is whether there were fewer assessments with data quality issues in the second year of implementation, which was considered by districts to be a high-stakes year at the time of test implementation. In the 2012-13 school year, assessments in 27 of the 54 district/course combinations (50%) contained data quality issues (see Table 4A). Of the 27 district/courses containing data quality issues, over half contained more than one data quality issue (e.g., bi-modal distribution and skew on the pretest). In 2013-14, assessments from 19 district/course combinations out of 59 (32%) displayed data quality issues; with the exception of District D, these assessments only contained one major issue (see Table 4B).

[insert Table 4A and 4B]

Districts appeared to remedy issues with many of the assessments in year one: of the 27 courses with data quality issues in year one that also contained an SLO in year two, 13 (48%) did not have data quality issues in the following year. However, eight of the 19 district/course combinations with data quality issues in year two did *not* contain data quality issues in the prior year (2012-13), which may be partly due to the introduction of new tests from year one to year two.

Overall, these results suggest that district modifications, particularly to the administration/grading procedures along with the match between the difficulty of the assessment and ability of students, led to an improvement in assessment data quality. It is unclear whether this was the natural result of districts addressing problems observed during the first year of implementation, or was the result of districts responding to pressure imposed from the intended high-stakes nature of results in year two. Moreover, it is unclear whether any improvements were the result of evidence suggested by item analyses, or general feedback from teachers and leaders about the quality and difficulty of the assessment.

Score manipulation

Score manipulation presents a second factor which can affect student SLO results. Score manipulation is likely to occur when (a) scores from an assessment are used in a high-stakes manner; (b) those grading the assessment have knowledge of the cut score used in the high-stakes decision; and (c) those grading the assessment are directly affected by the final score (Liguori, 2011; Koretz, 2008). Thus, there is reason for concern with SLO assessments in this state since student scores on SLO assessments from the 2013-14 school year were intended to be used for high-stakes decisions involving teachers at the end of the 2014-15 school year, and teachers were not only aware of the target score each student needed to reach but typically graded each student's pretest and posttest.[23]

There are two primary ways that score manipulation might occur within this state. The first is if teachers grade pre-assessments more stringently, thereby depressing

[23] In every district, either the teacher, school principal, or district calculates each student's target score. If the district calculates target scores, teachers are supplied with the target formula.

prescores and corresponding target scores, which will make it easier for students to achieve their target score. The second is if teachers inflate student postscores, particularly for those students with actual scores just below their target score, which would increase the number of students who meet their target score.

Determining if score manipulation is occurring under the first scenario- i.e., depressing student prescores – would require item-level data or an alternative measure of each student's achievement at baseline with which to compare SLO prescores, neither of which I have. Simply examining whether pretest scores rose from year one (a non-high stakes year) to year two (a high-stakes year) does not provide sufficient evidence, given the confounding factors of changes in assessments, targets and student cohort.

To determine if the latter scenario is occurring – adjusting student postscores for those just below their target – I follow the work of Dee, Jacob and McCrary (2011) in using a test of frequencies of postscores near the target score. Specifically, I compare frequencies of postscores near the target score in a high-stakes year (2013-14), in which one might expect to see score manipulation, to a low-stakes year (2012-13), in which one would not expected to see score manipulation.

I limit the sample to those district/courses in which teachers grade their own assessments. This included teachers from every course in Districts A, B, and G, as well as teachers in District H that taught British Literature and Literature and Comprehension.[24] I further limited the sample to those teachers who teach the same course in both the high-stakes and low-stakes school year, in order to compare frequencies for the same teachers

[24] District H uses automated scoring for the multiple choice items, and teacher scoring for the open-ended items. Open-ended items are only included in British Literature and Literature and Comprehension. Districts D and E used mostly automated scoring in year two, and it is unknown whether teachers graded their own assessments in District C and F.

in the same courses over time. Finally, I removed any district/course in which less than 200 students took the assessment. I am left with student scores from 660 teachers in 22 district/course combinations in each year (see Appendix B for the frequency of student scores included in the analysis in each year).

Recall that the variable of interest is *post_target*, which is equal to each student's postscore minus his/her target score. I compared the density of *post_target* scores that occur close to zero (i.e., the breakpoint) in year one to year two, under the assumption that one would expect to see manipulation of scores in a high-stakes year (year two) only. To do so, I first calculated McCrary log density estimates by district/course for each year. Next, I calculated a z-ratio by taking the difference in estimates of a given district/course in each year and dividing it by the standard error of the difference.[25] If the z-ratio for a given district/course is greater than a critical value of 1.96, it indicates that the density of scores to the right of the breakpoint is significantly greater than the density of scores to the left of the breakpoint in 2013-14 relative to 2012-13, which suggests the presence of score manipulation in a high-stakes year.

A benefit of this method is that easier assessments and targets in year two will not necessarily bias my estimates as they would if I were simply comparing frequencies of students meeting their target to frequencies of those not meeting their target. While easier targets and assessments in year two will likely lead to *more* students meeting their target, I would expect this to translate into increased frequencies *across* the distribution of positive *post_target* values and decreased frequencies across the distribution of negative *post_target* values. I would not expect easier targets and assessment to lead only to a

[25] The standard deviation of the difference=$\sqrt{{se_{year1}}^2 + {se_{year2}}^2}$

jump in the frequency of scores for those who exactly met their target compared to those who just missed their target. Further, by determining whether there is a significant difference in the discontinuity between year two and year one, I am reducing the likelihood of reporting a finding of score manipulation that is actually due to chance, by differencing out any stable aspect over time in the administration procedures, scoring procedures, or assessment itself that may be creating a discontinuity for reasons other than manipulation.

In Table 5, I report, for each district/course in the sample, the McCrary log estimates for each year (i.e., the log difference of intercepts from localized regressions on either side of the breakpoint) (columns 1 and 2); the standard errors of the McCrary estimates (columns 1 and 2, in parentheses); the difference between the 2013-14 and 2012-13 McCrary estimates (column 3), the standard deviation of the difference estimate (column 3, in parentheses); and the z-ratio of the difference estimate (column 4). Note that log density estimates using the McCrary test cannot be calculated for District G's third grade ELA and third grade Mathematics course due to the fact that the post-assessment in both courses is out of a small number of points; therefore, there are too few values of *post_target* scores to the left and right of the breakpoint with which to fit a local linear regression model.

[insert Table 5]

Out of the 20 district/course combinations, McCrary log estimates from two district/courses were significantly higher in year two than year one: District A's third grade Mathematics assessment and District H's Literature and Comprehension assessment. Importantly, these results do not appear to be sensitive to changes in

bandwidth/bin-size. In Figure 5a, I display the estimated density function of the variable *post_target* in District A's grade 3 mathematics assessment in each year; in Figure 5b, I display the estimated density function of the variable *post_target* in District H's high school Literature and Composition assessment in each year. Scores from both district/courses clearly show evidence of a larger discontinuity at the breakpoint of 0 in 2013-14 (right hand panel of both figures) compared to 2012-13 (left hand panel of both figures).

[insert Figure 5]

While there is evidence consistent with score manipulation in two of the 20 districts/courses, given the issue of multiple comparisons, at least 1 in 20 estimates are expected to be significant simply by chance based on an alpha level of .05. To address this issue, I apply the Bonferroni correction and calculate a new alpha level of .00125 ($\alpha/n = .025/20$), which translates to a critical z-value of 3.08. Under this conservative z-test, estimates from both course do not continue to show significant evidence of score manipulation.

There is considerable concern among policymakers and journalists about the possibility of score manipulation in teacher evaluation systems in states implementing SLOs, given how SLOs are graded (Scott, 2013; Marion et al., 2012). Using the McCrary test as one mechanism for detecting score manipulation around the target score, I fail to find evidence that it is occurring. Informal conversations with district and state personnel suggest that teachers did not truly understand the high stakes nature of these tests, which may have been hampered by fact that 2013-14 scores are lagged, and will not be reported in teacher evaluations until the end of 2014-15. Some teachers, for example, failed to

record student prescores or postscores, suggesting that these teachers believed the assessment to be voluntary. Further, it is possible that teachers did not fully understand how student or teacher SLO scores were calculated.

Choice and calculation of student growth target

A student's SLO classification in this state is based on whether his/her postscore is equal to or greater than his/her target score. Target scores for each course can be calculated using one of four different target types, each of which can vary in terms of the rigor (difficulty) expected of students. Districts are required to set the target type for each course, as well as the rigor/difficulty of the target for each course. Teachers are typically expected to calculate each student's target score after calculating each student's prescore, since target scores based on each target formulation vary as a function of prescore. Given the way the student growth targets are developed and applied in this state, an examination of how changes to the type and rigor of the target impacts student SLO attainment, and an examination of whether errors in target calculations affect SLO attainment, is required.[26]

Choice of student growth target. Recall that there are four different types of target formulations used in this state. The first, an *individual growth target*, sets each student's required postscore as a function of a percentage gain from his/her prescore towards the maximum score on the posttest. The second, *a tiered growth target*, sets cut-points on the pretest and requires students scoring in each tier to improve on the posttest by a set number of points, with the number of required points decreasing for each successive tier. The third is a *uniform growth target*, whereby each student must increase from their score from the pretest to the posttest by the same number of points. And the

[26] In Appendix A, I provide a brief discussion of two additional issues: unintended consequences of two target types and the calculation of target scores when the prescore is missing.

final target type is a *categorical growth target*, whereby cut-scores are imposed on the pretest and posttest, and students falling in one category on the pretest are expected to reach the next successive category on the posttest.

A summary of the target types used by districts in the sample is provided in Table 6. Evident, first, is that the individual target type is the most common target type, every target formulation was used across districts in the sample. Second, two districts modified their target types from year one to year two: (1) District B switched from a categorical target to an individual target in third grade ELA and Mathematics; and (2) District D switched from an individual target to a uniform target for third grade ELA and Mathematics.[27] While District E did not alter their target type for the existing SLO courses, they did employ an individual target in two new SLO courses.

[insert Table 6]

The primary change to student targets that districts enacted from year one to year two was to reduce the level of *rigor* required of the target formulation. In year one, for example, many districts required an individual target whereby students were expected to grow by 50% or even 65% of their potential growth. In year two, districts typically lowered the percentage gain required to between 20% and 33%. Similar reductions in rigor appear to have been made with the other target types.

Effect of changes in target type. The graph in Figure 6 illustrates the relationship among each alternative target type. Each line shows the target score, defined as the minimum postscore necessary for a given prescore under a particular target formulation. Evident is that when the level of rigor is held constant, the relationship between the

[27] It is likely that both District B and District D changed their target types because they changed their assessments for these two courses, as noted in the previous section on assessment data quality.

required postscore and prescore, in general, looks very similar among the individual target, uniform target and tiered target. However, the categorical target shows the greatest divergence from the other three target types that I defined.

[insert Figure 6]

The extent to which teacher scores differ across alternative target types will be a function of where student pretest and posttest scores fall in Figure 6. In general, if student pre- and postscores fall above or below the lines, then student results will be the same across target type (i.e., regardless of target type, student postscores will be equal to, greater than or less than their target score). If student scores fall in between the lines in Figure 6, then the student results will differ depending on target type. For example, if a student scores a 10 on the pretest and a 30 on the posttest, he/she will meet the target score under the individual, tiered and uniform target that I defined, but not the categorical target. On the other end of the prescore distribution, if a student scores a 90 on the pretest and a 90 on the posttest, he/she will meet the required target score under the categorical target, but not under the remaining three target types.

What happens when alternative target types are applied to the underlying student data? Table 7 provides the average percentage of students meeting each alternative target score. Evident is that holding the rigor of the target constant, the choice of target type does in fact alter student performance, by as much as 12 percentage points in District B and as little as 2 percentage points in District A.

[insert Table 7]

What Table 7 does not show is the extent to which student rankings will change as a result of shifting target types. Therefore, Table 8 displays the relationship among

alternative target scores based on Spearman rank correlations. Evident is that correlations are nearly perfect among the individual, tiered and uniform targets, suggesting that student target scores rank consistently across these three target types. However, student target scores may change rank if a categorical target is applied instead of one of the other three target types.

[insert Table 8]

Effect of shifts in target rigor. In addition to changes in target type, it is important to understand how differences in the level of rigor of the target can affect the number of students who are designated as having met their SLO growth target. Table 9 displays the district-level percentage of students meeting each alternative *individual* growth target, whereby the rigor of the target varies but the target type is held constant ($Target_{Indiv_{25}}$, $Target_{Indiv_{50}}$, $Target_{Indiv_{60}}$). Changes to the rigor of the growth target shift student SLO attainment within districts by as little as 19 percentage points in District E and as much as 51 percentage points in District F. Evident from this simulation is that variation in target rigor has a greater effect on student SLO scores compared to variation in the type of target.

[insert Table 9]

While changing target rigor will affect the percentage of students meeting their target, it will do so uniformly, by shifting the standard up or down. As a result, changing the rigor of the target within a district/course will not affect rankings of students within a course; however it will affect the comparability of scores across courses within a district or across districts if the same standard is not applied uniformly to all students.

Target errors. Every student within the same district/course should be assigned the same target formulation, according to state implementation rules. Therefore, all students within a course with the same prescore are expected to have the same target score regardless of student, classroom, or school-level demographics. In this state, target scores are typically calculated by the classroom teacher instead of calculated in an automated fashion. Therefore, an additional concern is the extent to which target scores are being calculated accurately and the effect of potential errors on overall student SLO attainment.

Target errors in year one. In the first year of implementation, target errors occurred in five of the eight districts in the sample, four in which teachers calculated student target scores. Notably, target errors were not found in District D and District E, where target scores were primarily automated.

Errors can be attributed to both rounding discrepancies and incorrect application of the target, and occur with every target type. In 2012-13, a total of 1.9% of students in the sample (n=2,231) had an incorrect target score. Of students with target errors, 16% (n=353) had incorrect SLO attainment classifications (i.e., were incorrectly determined to have met or not met their target score). Note that students were just as likely to have been classified incorrectly with a "did not meet SLO" designation as they were to have been classified incorrectly with a "met SLO" designation, suggesting that these errors were simply innocuous mistakes rather than intentional. The reasons for target errors in each district are discussed in greater detail below.

In Districts A and H, which used individualized target formulations, target errors appear to be the result of rounding discrepancies in the calculation of the target

formulation. For example, a student who should have a target score of 70.15 was instead assigned a target score of 70.1 or 70.2.[28] In Districts B and G, students who took the Mathematics Grade 3 assessment, in which a tiered and categorical target, respectively, were applied, a small number of students with a given prescore were given a target score based on an easier or harder to reach tier/category. For example, on District G's third grade mathematics assessment, approximately 36% of students with a prescore of 4 were given a target score of 6 instead of a target score of 5. Finally, in District C, which used an individualized target formulation in every subject, target errors were due to two issues: first, some teachers applied a lower percentage gain for some students (30% instead of 50%).[29] Second, the district required a cap on their target scores, such that students were not given a target score greater than 90 points on a 100 point assessment; however in certain instances, this cap was not applied.

Target errors in year two. Target errors continued to occur in the second year of implementation, due solely to rounding discrepancies under the individual target.[30] However, target errors were far less frequent compared to year one: less than 1% (n=288) of all students in the 2013-14 sample were given the wrong target score. Of students with target errors, 30% (n=85) had an incorrect SLO attainment classification.[31] Table 10 displays the number of target errors, by district and year.

---

[28] Conversations with District A personnel suggest that these errors are indeed the result of human error.
[29] This type of misclassification in District C occurs across many different teachers 'classrooms but not always for all students within a classroom. While it is possible that the district may have given teachers guidance to reduce the percentage gain for students with certain characteristics, this reduction in percentage gain does not correlate with student-level variables included in my dataset (i.e., ELL, SWD).
[30] Informal conversations with district personnel suggest that target calculations were often standardized at the district level in year two of implementation, which likely contributed to the decrease in target misclassification.
[31] I exclude cases in District D where students in a given course have a prescore of 0 but different target scores, based on conversations with district personnel suggesting that that 0 was used as a placeholder for missing data.

[insert Table 10]

Discussion

Given the lack of research on the quality of assessment data and selection of student growth targets used to do determine student SLO attainment, each aspect requires careful consideration in order to ensure that inferences from student SLO results are valid. My investigation into the quality of assessment data suggests serious issues with some of the assessments being administered for SLOs, including inconsistent test administration and scoring procedures among teachers administering the same SLO, which can place scores of some students on a different scale than others taking the same assessment. I also found evidence of mismatch between the difficulty of the assessment and the ability of students, along with a lack relationship between student pre and postscores, both of which can distort student growth score interpretations. While visual examinations of the scale score distributions and descriptive statistics do not provide definitive evidence of low quality assessments, they do suggest that the assessment data may not be as reliable as data from state standardized assessments. This is particularly true since, given the need to implement SLOs quickly under federal policies, the tests likely lack evidence of internal consistency reliability, and evidence validating their use in a high-stakes evaluation system. Therefore, it is important for states and districts implementing SLOs to investigate the quality of their assessments, prior to the use of the results in a high-stakes framework. To this end, several authors, including Steele et al. (2011), Lachlan-Hache et al., (2012) and Herman et al. (2011) offer concrete guidelines

for developing and improving the validity and reliability of assessments used in non-tested grades and subjects.

While assessments administered in both years displayed serious issues which could affect the interpretation of student growth scores, the number of assessments with data quality issues declined considerably from the first to the second year of implementation. In particular, fewer assessments in year two contained anomalous distributions, suggesting that districts made changes to ensure consistent test administration/grading and to ensure that assessments were set at the appropriate level of difficultly. On the other hand, the relationship between student prescores and postscores remains low for many assessments in year two, which could be the result of actions taken in an intended high-stakes year to improve student results on the postscore.

Furthermore, the accuracy with which student target scores are calculated needs to be examined, in light of the fact that target score calculations are typically calculated by teachers using an excel spreadsheet, and errors due to both inconsistent rounding and inconsistent application of the target formulation occur. In fact, two percent of students in year one had incorrect target scores, a number that substantially declined by year two. Based on an understanding of changes in district processes over time, this reduction in errors is likely due to greater standardization of the procedures for calculating student growth targets.

Despite the fact that teachers are grading the assessments that will be used in their own evaluations, there was no evidence based on the McCrary density test to suggest score manipulation. It is possible that score manipulation is in fact occurring, but that I was not able to detect it because I was looking for a very specific type: inflation of scores

right near the target score. However, informal conversations with state and district personnel suggest that teachers were not necessarily aware that the SLO scores from the 13-14 school-year would be incorporated into evaluation scores in the following school year. Given the lagged nature of scores, I hypothesize that evidence of score manipulation will not appear until after teachers have had a chance to see their evaluation scores and fully understand the system used to calculate their scores. That being said, if most districts switch to automated or district-level scoring, which some are beginning to do, I would not expect to find any evidence of SLO score manipulation in this state.

Moreover, additional investigations into the extent to which other forms of score distortion are occurring, for example teaching to the test and a narrowing of the curriculum, are still needed (Steele, et al., 2011). Informal conversations with district and school leaders indicate that this is a concern, and low pre-postscore correlations coupled with high gain scores suggest this could be the case. Such actions could serve to bias teacher scores if students in certain courses are doing better on the posttest simply due to test preparation, for example, rather than actual growth in learning (Koretz, 2008).

In addition to data quality issues with both assessment scores and target scores, I investigated how both the type and rigor of the chosen growth target will impact inferences of student performance. Similar to research on the effect of alternative growth models on student performance and school attainment under NCLB (Hoffer et al., 2011), shifts in the type and especially rigor of the growth target will have a substantial impact on the classification of student SLO attainment. This has implications for the comparability of scores both within and across districts, since target type and rigor varies within districts across courses and over time, as well as across districts.

This is not to say that targets should be standardized, since districts must consider first which type of growth target best meets their policy goals. For example, if the district wishes to ensure that the target formulation chosen does not place a greater burden on lower-performing students, then a growth formula that requires all students to grow by the same absolute amount, such as a uniform target, is preferable. On the other hand, if the district wants to ensure that lower-performing students demonstrate greater growth relative to higher-performing students, a growth formula which requires increasing growth as a function of decreasing prescores, such as with an individualized target requiring a percentage gain based on closing the distance between one's prescore and total number of points on the test, is preferable.

Further, standardizing targets without standardizing the assessments will continue to lead to lack of comparability of scores since harder assessments will require easier targets in order to ensure that students can still meet their growth target. One recommendation for district or school personnel in charge of determining the target formulation would be to introduce a standard-setting procedure whereby target scores are tied concretely to pre-defined criteria and prior student data is used to determine attainment feasibility rather than simply applying a formula to existing student prescore data.

Tables and Figures

Table 2. Frequency of students in 2012-13 sample (first row) and 2013-14 sample (second row), by district/course.

	ELA grade 3	Math grade 3	British Literature	Literature & Comprehension	World Literature	Chemistry	Algebra	American Government	World History	Total
District A		2,440	1,079	1,481		979	895	1,646	1,630	10,150
		2,202	528	1,004		947	1,077	1,170	1,353	8,281
District B	3,284	3,941	105	2,479		1,652		69	2,095	13,625
	3,273	3,697	1,945	3,265		2,074		2,027	2,572	18,853
District C	3,851	4,157	1,971		2,953	2,294	2,959	2,906	3,484	24,575
	4,041	4,249	1,601		3,075	1,632	2,398	2,850	3,269	23,115
District D	5,247	4,925	2,779		3,555	3,989	3,230	2,677	3,000	29,402
	7,937	8,647	4,114	2	4,435	6,464	4,855	6,778	5,413	48,645
District E		639	133	608		37		116	102	1,635
	3,940	99	1,791	3,161		7,319	2,044	1,515	4,457	24,326
District F	2,704	2,426	838		2,249	1,038	1,368	4,097	575	15,295
	274	1,888	124		171	261	354	522	660	4,254
District G	1,987	2,421	1,319	1,876		1,136	1,568	463	1,851	12,621
	2,736	2,393	1,702	2,789		1,260	2,109	1831	2,207	17,027
District H	2,048	2,166	953	916	73	661	1,071	1,166	1,477	10,531
	2,164	2,126	1,536	1,710	375	1,071	27	2,422	1,668	13,099

Notes: The first row of each district is the number of students in 2012-13 included in the sample, the second row of each district is the number of students in 2013-14 included in the sample.

Table 3. Student growth target formulations for each alterantive target type.

Target Type	Target Formula Applied
Individual	$Target_i^{individual} = x_{prescore} + .23\left(x_{\max score} - x_{prescore}\right)$
Uniform	$Target_i^{uniform} = x_{prescore} + .16(x_{maxscore}), \lim_{n \to x_{maxscore}} x_{maxscore}$
Tiered	$Target_i^{tier} = x_{prescore+}$ $\begin{cases} .20x_{\max score}, & x_{prescore} \leq .30x_{maxscore} \\ .15x_{\max score}, & .3x_{maxscore} < x_{prescore} \leq .5x_{maxscore} \\ .10x_{\max score}, & .5x_{maxscore} < x_{prescore} \leq .75x_{maxscore} \\ .05x_{\max score}, & x_{prescore} > .75x_{maxscore,} \lim_{n \to x_{maxscore}} x_{maxscore} \end{cases}$
Categorical	$Target_i^{categorical} =$ $\begin{cases} .31x_{\max score}, & x_{prescore} \leq .30x_{maxscore} \\ .56x_{\max score}, & .30x_{maxscore} < x_{prescore} \leq .55x_{maxscore} \\ .76x_{\max score}, & .55x_{maxscore} < x_{prescore} \leq .75x_{maxscore} \\ .91x_{maxscore}, & .75x_{maxscore} > x_{prescore} \end{cases}$

Notes. The tiered target and uniform target formulations have limits, such that a student will not have target score greater than the maximum number of points on an assessmnet. The rigor of each target formulation has been constructed so that the same number of students in the sample meet their target under each alternative target formulation.

Table 4A. Summary of 2012-13 assessment data quality, by district/course.

District/ Course	Course included analysis	Bi-modal prescores	Bi-modal postscores	Illegal prescores	Illegal postscores	Pretest skew	Posttest skew	Pretest spike	Posttest spike	Low pre-post corr	Data Quality Issues
District A ELA Gr3	no										
District A Brit Lit	yes	No	No	yes	No	no (0.11)	no (-0.38)	No	No	no (0.52)	No
District A Lit/Comp	yes	No	No	No	No	no (-0.17)	no (-0.57)	No	No	no (0.52)	No
District A World Lit	no										
District A Chem I	yes	No	No	No	yes	no (0.35)	no (0.52)	No	No	no (0.63)	No
District A Math Gr3	yes	No	No	No	yes	no (-0.43)	yes (4.66)	No	No	no (0.69)	No (skew is driven by one illegal score)
District A Algebra	yes	No	No	No	No	no (0.56)	no (0.34)	yes	no	yes (-0.19)	Yes
District A Am Govt	yes	No	No	No	No	no (0.33)	no (-0.21)	No	No	no (0.21)	No
District A History	yes	No	No	No	No	yes (2.18)	no (0.30)	No	No	no (0.38)	Yes
District B ELA Gr 3	yes	No	No	No	No	no (1.11)	no (0.97)	No	No	no (0.71)	No
District B British Lit	no										
District B Lit/Comp	yes	yes	yes	No	No	yes (2.16)	no (1.95)	No	No	no (0.88)	Yes
District B World Lit	no										
District B Chem I	yes	No	No	No	No	no (0.77)	no (-0.45)	No	No	no (0.37)	No

District B Math Gr3	yes	yes	yes	No	yes	no (-0.36)	no (-0.01)	no	no	no (0.87)	Yes
District B Algebra	no										
District B Am Govt	no										
District B History	yes	No	No	No	No	no (0.28)	no (-0.18)	No	No	no (0.38)	No
District C ELA Gr3	yes	No	No	yes	No	no (1.14)	no (0.53)	yes	yes	no (0.37)	Yes
District C British Lit	yes	No	No	No	No	no (0.24)	no (-0.13)	yes	yes	yes (0.07)	Yes
District C Lit/Comp	no										
District C World Lit	yes	No	No	No	No	no (-0.17)	no (0.10)	yes	yes	yes (0.17)	Yes
District C Chem I	yes	No	No	No	yes	no (-0.25)	no (0.40)	yes	yes	yes (0.16)	Yes
District C Math Gr3	yes	No	No	yes	No	no (1.28)	no (0.28)	yes	yes	no (0.39)	Yes
District C Algebra	yes	No	No	No	No	no (1.79)	no (0.96)	yes	yes	yes (0.16)	Yes
District C Am Govt	yes	No	No	No	No	no (0.33)	no (0.49)	yes	yes	yes (0.00)	Yes
District C History	yes	No	No	No	No	yes (3.58)	no (0.61)	yes	yes	yes (0.09)	Yes
District D ELA Gr3	yes	No	No	No	No	no (1.03)	no (0.17)	yes	no	no (0.44)	Yes
District D British Lit	yes	No	yes	No	No	no (0.60)	no (0.70)	yes	no	yes (0.17)	Yes
District D Lit/Comp	no										
District D World Lit	yes	No	yes	No	No	no (0.88)	no (0.56)	yes	No	yes (0.12)	Yes
District D Chem I	yes	No	No	No	No	no (0.16)	no (0.44)	yes	No	no (0.24)	Yes

District D Math Gr3	yes	No	No	No	No	no (1.16)	no (0.10)	yes	No	no (0.51)	Yes
District D Algebra	yes	No	No	No	No	no (0.52)	no (0.99)	yes	No	no (0.27)	Yes
District D Am Govt	yes	No	No	No	No	no (1.17)	no (0.68)	yes	No	no (0.25)	Yes
District D History	yes	No	No	No	No	no (1.21)	no (0.61)	yes	No	yes (0.13)	Yes
District E ELA Gr3	no										
District E British Lit	no										
District E Lit/Comp	yes	No	No	No	No	no (0.16)	no (0.58)	No	No	no (0.57)	No
District E World Lit	no										
District E Chem I	no										
District E Math Gr3	yes	yes	yes	No	No	no (-0.26)	no (-0.83)	No	No	no (0.65)	Yes
District E Algebra	no										
District E Am Govt	no										
District E History	no										
District F ELA Gr3	yes	No	No	No	No	no (0.42)	no (-0.26)	No	No	no (0.36)	No
District F British Lit	yes	yes	yes	yes	No	no (-0.28)	no (-1.19)	No	No	no (0.77)	Yes
District F Lit/Comp	no										
District F World Lit	yes	yes	yes	No	No	no (-0.22)	no (-0.66)	No	No	no (0.85)	Yes
District F Chem I	yes	No	No	No	No	no (0.24)	no (-0.36)	No	No	no (0.48)	No

District F Math Gr3	yes	No	No	No	No	yes (10.46)	no (0.13)	yes	No	yes (0.09)	Yes
District F Algebra	yes	No	No	No	No	no (0.43)	no (-.17)	No	No	no (0.23)	No
District F Am Govt	yes	No	No	No	No	no (0.48)	no (0.50)	No	No	no (0.42)	No
District F History	yes	No	No	No	No	no (0.96)	no (0.24)	No	No	no (0.56)	No
District G ELA Gr3	yes	No	No	No	No	no (0.70)	no (-0.25)	No	No	no (0.46)	No
District G British Lit	yes	No	No	No	No	no (0.42)	no (-0.50)	No	No	no (0.41)	No
District G Lit/Comp	yes	No	No	No	No	no (0.04)	no (-0.61)	No	No	no (0.42)	No
District G World Lit	no										
District G Chem I	yes	No	No	No	No	no (-0.23)	no (0.31)	No	No	no (0.56)	No
District G Math Gr3	yes	No	No	No	No	no (-1.00)	no (-0.18)	No	No	no (0.45)	No
District G Algebra	yes	No	No	No	No	yes (5.90)	no (0.38)	No	No	no (0.31)	Yes
District G Am Govt	yes	No	No	No	No	no (1.73)	no (0.09)	No	No	no (0.46)	No
District G History	yes	No	No	No	No	yes (2.14)	no (0.15)	No	No	no (0.24)	Yes
District H ELA Gr3	yes	No	No	No	No	no (0.21)	no (-0.51)	No	No	no (0.67)	No
District H British Lit	yes	No	No	No	No	no (-0.24)	no (-0.94)	No	No	no (0.59)	No
District H Lit/Comp	yes	No	No	No	No	no (-0.32)	no (-0.61)	No	No	no (0.61)	No
District H World Lit	no										
District H Chem I	yes	No	No	No	No	no (0.47)	no (-0.09)	No	No	no (0.71)	No

District H Math Gr3	yes	No	No	No	No	no (0.83)	no (-0.12)	No	No	no (0.63)	No
District H Algebra	yes	No	No	No	No	no (0.75)	no (-.02)	No	No	no (0.38)	No
District H Am Govt	yes	No	No	No	No	no (0.36)	no (-0.19)	No	No	no (0.54)	No
District H History	yes	No	No	No	No	yes (2.73)	no (0.37)	No	No	no (0.40)	Yes
72 courses	54 courses	5 pretests	7 posttests	4 pretests	4 posttests	7 pretests	1 posttest	18 pretests	8 posttests	11 district/ courses	27 district/ courses

Table 4B. Summary of 2013-14 assessment data quality, by district/course.

District/ Course	Course included analysis	Bi-modal prescores	Bi-modal postscores	Illegal prescores	Illegal postscores	Pretest skew	Posttest skew	Pretest spike	Posttest spike	Low pre-post corr	Data Quality Issues
District A ELA Gr3	No					.	.				
District A Brit Lit	Yes	no	no	no	no	no (0.93)	no (-0.40)	no	no	no (0.39)	no
District A Lit/Comp	Yes	no	no	no	no	no (0.20)	no (-0.28)	no	no	no (0.27)	no
District A World Lit	No					.	.				
District A Chem I	Yes	no	no	no	yes	no (0.16)	no (-0.20)	no	no	yes (0.09)	yes
District A Math Gr3	Yes	no	no	no	yes	no (-0.28)	no (-0.32)	no	no	no (0.48)	no
District A Algebra	Yes	no	no	no	no	yes (3.84)	no (0.85)	yes	no	yes (0.01)	yes
District A Am Govt	Yes	no	no	no	no	no (0.38)	no (0.12)	no	yes	no (0.31)	yes
District A History	Yes	no	no	no	no	no (0.44)	no (0.12)	no	yes	yes (0.09)	yes
District B ELA Gr 3	Yes	no	no	no	no	no (-0.02)	no (-0.75)	no		no (0.67)	no
District B British Lit	Yes	no	no	no	no	no (-0.13)	no (-1.26)	no		no (0.53)	
District B Lit/Comp	Yes	no	no	no	no	no (-0.30)	no (-1.06)	no		no (0.64)	no
District B World Lit	No					.	.				
District B Chem I	Yes	no	no	no	no	no (0.70)	no (-0.54)	no		no (0.39)	no
District B Math Gr3	Yes	no	no	no	no	no (0.52)	no (-0.39)	no		no (0.62)	no

District B Algebra	No					.	.				
District B Am Govt	Yes	no	no	no	no	no (0.38)	no (-0.52)	no		no (0.49)	no
District B History	Yes	no	no	no	no	no (0.21)	no (-0.14)	no		no (0.54)	no
District C ELA Gr3	Yes	no	no	no	no	no (0.75)	no (0.24)	no		no (0.68)	no
District C British Lit	Yes	no	no	no	no	no (-0.08)	no (-0.36)	no		no (0.29)	no
District C Lit/Comp	No					.	.				
District C World Lit	Yes	no	no	no	no	no (-0.09)	no (-0.32)	no	no	no (0.48)	no
District C Chem I	Yes	no	no	no	no	no (0.38)	no (0.42)	no	no	no (0.26)	no
District C Math Gr3	Yes	no	no	no	no	no (0.74)	no (0.25)	no	no	no (0.64)	no
District C Algebra	Yes	no	no	no	no	no (0.62)	no (0.84)	no	no	yes (0.12)	yes
District C Am Govt	Yes	no	no	no	no	no (0.38)	no (0.55)	no	no	no (0.32	no
District C History	Yes	no	no	no	no	no (0.71)	no (0.37)	no	no	no (0.49)	no
District D ELA Gr3	Yes	yes	yes	no	no	no (0.83)	no (0.67)	no	no	no (0.86)	yes
District D British Lit	Yes	yes	yes	yes		no (0.28)	no (-0.45)	Yes	yes	Yes (0.16)	yes
District D Lit/Comp	No					.	.				
District D World Lit	Yes	no	no	no	no	no (0.28)	no (.06)	yes	yes	Yes (0.02)	yes
District D Chem I	Yes	no	no	no	no	no (0.57)	no (0.47)	yes	yes	yes (0.19)	yes
District D Math Gr3	Yes	no	no	no	no	no (-0.56)	no (-0.67)	No	no	no (0.79)	no

District D Algebra	Yes	no	no	no	no	no (0.53)	no (0.37)	yes	yes	yes (-0.03)	yes
District D Am Govt	Yes	no	no	no	no	no (0.48)	no (0.13)	yes	yes	yes (0.01)	yes
District D History	Yes	no	no	no	no	no (0.65)	no (-0.04)	yes	yes	Yes (0.07)	yes
District E ELA Gr3	Yes	no	no	no	no	no (0.01)	no (-0.77)	no	no	no (0.66)	no
District E British Lit	Yes	no	no	no	no	no (-0.15)	no (-0.32)	no	no	no (0.46)	no
District E Lit/Comp	Yes	no	no	no	no	no (0.46)	no (-0.05)	no	no	no (0.60)	no
District E World Lit	No					.	.				
District E Chem I	Yes	no	no	no	no	no (0.37)	no (-0.34)	no	no	no (0.51)	no
District E Math Gr3	No					.	.				
District E Algebra	Yes	no	no	no	no	no (0.28)	no (0.08)	no	no	no (0.20)	no
District E Am Govt	Yes	no	no	yes	yes	no (0.135)	no (0.07)	no	no	no (0.46)	no
District E History	Yes	no	no	yes	yes	no (0.00)	no (0.26)	no	no	no (0.49)	no
District F ELA Gr3	Yes	no	no	no	no	no (1.49)	no (-0.68)	no	no	yes (0.14)	yes
District F British Lit	No										
District F Lit/Comp	No					.	.				
District F World Lit	No										
District F Chem I	Yes	no	no	no	no	no (-0.21)	no (-0.22)	no	no	no (0.43)	no
District F Math Gr3	Yes	no	no	no	no	yes (4.82) but driven	no (0.16)	no	no	no (0.24)	No

						by low number of scores on test					
District F Algebra	Yes	no	no	no	no	no (-0.16)	no (0.19)	no	no	yes (-0.08)	yes
District F Am Govt	Yes	no	no	no	no	no (-0.10)	no (0.26)	no	no	no (0.26)	no
District F History	Yes	no	no	no	no	no (0.46)	no (-0.09)	no	no	no (0.36)	no
District G ELA Gr3	Yes	no	no	no	no	no (-0.36)	no (-0.70)	no	no	no (0.90)	no
District G British Lit	Yes	no	no	no	no	no (0.65)	no (-0.29)	no	no	no (0.36)	no
District G Lit/Comp	Yes	no	no	no	no	no (0.30)	no (-0.68)	no	no	no (0.37)	no
District G World Lit	No					.	.				
District G Chem I	Yes	yes	no	no	no	no (-0.02)	no (0.24)	no	no	yes (-0.05)	yes
District G Math Gr3	Yes	no	no	no	yes	no (-0.85)	yes (2.27) but driven by illegal scores	no	no	no (0.23)	no
District G Algebra	Yes	yes	no	no	no	no (0.12)	no (-0.41)	no	no	no (0.32)	yes
District G Am Govt	Yes	no	no	no	no	yes (2.94)	no (0.26)	no	no	no (0.36)	yes
District G History	yes	No	No	No	No	No (0.15)	No (-.02)	No	No	No (2.6)	no
District H ELA Gr3	Yes	no	no	no	no	no (-0.43)	no (-1.71)	no	no	no (0.69)	no
District H British Lit	Yes	no	no	no	no	no (0.39)	no (0.18)	no	no	no (0.36)	no
District H Lit/Comp	Yes	no	no	no	no	no (0.183)	no (-0.32)	no	no	no (0.49)	no

District H World Lit	Yes	no	no	no	no	no (-0.11)	no (-0.92)	no	no	yes (0.04)	yes
District H Chem I	Yes	no	no	no	no	no (0.34)	no (0.08)	no	no	yes (0.17)	yes
District H Math Gr3	Yes	no	no	no	no	no (-0.12)	no (-0.96)	no	no	no (0.59)	no
District H Algebra	No										
District H Am Govt	Yes	no	no	no	no	no (0.64)	no (0.33)	no	no	no (0.47)	no
District H History	Yes	no	no	no	no	no (0.60)	no (-0.40)	no	no	no (0.38)	no
	59 courses	4 pretests with bi-modal distribution	2 posttests with bi-modal distribution	3 pretests	5 posttests	2 pretests (and 1 driven by low number of points on test)	1 posttest (but driven by illegal scores)	6 pretests	8 posttests	15 district/ courses	19 district/ courses

Table 5. McCrary density score manipulation results, by district/course.

		2013-14 Log Estimate	2012-13 Log Estimate	Difference In Log Estimates	z-ratio
District A	Brit Literature	-0.212	-0.053	-0.159	-0.438
		(0.299)	(0.207)	(0.363)	
	Lit and Comp	0.524	0.101	0.423	1.733
		(0.138)	(0.201)	(0.244)	
	Chemistry	0.219	0.005	0.214	0.485
		(0.200)	(0.393)	(0.441)	
	Math grade 3	0.286	-0.089	0.375	2.249
		(0.121)	(0.115)	(0.167)	
	Algebra	0.270	0.149	0.121	0.280
		(0.252)	(0.351)	(0.432)	
	Am Gov't	0.263	0.796	-0.533	-1.483
		(0.227)	(0.278)	(0.359)	
	World History	-0.367	0.243	-0.610	-2.283
		(0.207)	(0.169)	(0.267)	
District B	ELA Grade 3	-0.153	0.261	-0.414	-2.649
		(0.115)	(0.107)	(0.156)	
	Lit and Comp	-0.011	-0.264	0.253	0.870
		(0.089)	(0.277)	(0.291)	
	Chemistry	-0.291	0.006	-0.297	-1.163
		(0.189)	(0.171)	(0.255)	
	Math Grade 3	0.117	0.127	-0.010	-0.065
		(0.114)	(0.096)	(0.150)	
	World History	-0.061	0.062	-0.122	-0.741
		(0.130)	(0.102)	(0.165)	
District G	Brit Literature	0.079	-0.212	0.291	1.160
		(0.185)	(0.170)	(0.251)	
	Lit and Comp	-0.091	0.493	-0.584	-2.658
		(0.172)	(0.136)	(0.220)	
	Chemistry	0.265	-0.044	0.308	1.411
		(0.162)	(0.147)	(0.218)	
	Algebra	0.082	0.268	-0.187	-0.682
		(0.186)	(0.201)	(0.274)	
	Am Gov't	0.224	0.572	-0.348	-0.701
		(0.299)	(0.396)	(0.496)	
	World History	-0.134	-0.322	0.188	0.882
		(0.123)	(0.174)	(0.213)	
District H	Brit Literature	0.125	-0.074	0.199	0.608
		(0.242)	(0.220)	(0.327)	

		2013-14 Log Estimate	2012-13 Log Estimate	Difference In Log Estimates	z-ratio
	Lit and Comp	0.164	-1.380	1.544	2.337
		(0.156)	(0.642)	(0.661)	

Notes. Standard errors (for columns 1 and 2) and the standard deviation (for column 3) are in parentheses.
Z-ratios are calculated by dividing the difference estimate by the standard deviation.

Table 6. Target type, by district.

	2012-13	2013-14
District A	Individual (7)	Individual (7)
District B	Individual (3), Category (2)	Individual (7)
District C	Individual (8)	Individual (8)
District D	Individual (8)	Individual (6), Uniform (2)
District E	Tier (2)	Tier (5), Individual (2)
District F	Individual (6), Uniform (2)	Individual (4), Uniform (2)
District G	Categorical (8)	Categorical (8)
District H	Individual (8)	Individual (8)
Total	Individual (40), Categorical (10), Tier (2), Uniform (2)	Individual (42), Categorical (8), Tier (5), Uniform (4)

Notes. Numbers in parentheses are the number of courses with the given target type. Only courses in the sample with student n-sizes less than 200 are included.

Table 7. Percentage of students in 2012-13 sample meeting each alternative target formulation, by district.

	Individual	Tiered	Uniform	Categorical	Maximum divergence by target type
District A	48%	47%	47%	50%	2%
District B	47%	45%	41%	53%	12%
District C	35%	36%	38%	35%	3%
District D	61%	64%	66%	56%	10%
District E	39%	38%	33%	40%	7%
District F	60%	59%	59%	66%	7%
District G	56%	52%	48%	52%	8%
District H	64%	63%	62%	64%	3%

Notes. N= 116,580. All students in courses with adequate sample sizes and with postscores less than or equal the maximum number of points on the test are included in the analyses.

Table 8. Spearman rank correlations of student target scores based on each alternative target type.

	Individual	Tiered	Uniform	Categorical
Individual	1.000			
Tiered	0.996	1.000		
Uniform	0.994	0.996	1.000	
Categorical	0.890	0.869	0.880	1.000

Notes. N= 116,580. All students in courses with adequate sample sizes and with postscores less than or equal the maximum number of points on the test are included in the analyses.

Table 9. Percentage of students in 2012-13 sample meeting alternative individual student growth targets that shift in rigor, by district.

	25% growth	50% growth	60% growth	Change in percentage growth when target rigor shifts from 25% to 60%
District A	46%	21%	12%	34%
District B	46%	22%	14%	32%
District C	34%	12%	7%	27%
District D	58%	25%	16%	42%
District E	39%	26%	20%	19%
District F	58%	20%	8%	51%
District G	55%	25%	14%	41%
District H	61%	26%	15%	46%

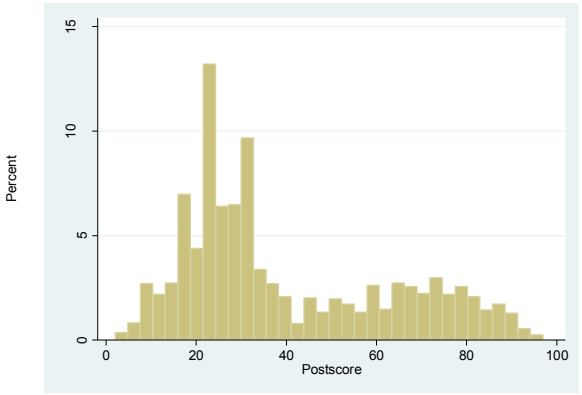
Notes. N= 116,580. All students in courses with adequate sample sizes and with postscores less than or equal the maximum number of points on the test are included in the analyses.

Table 10. Frequency of students with target errors, by district and year.

	2012-13	2013-14
District A	309	144
District B	616	35
District C	407	14
District D	-	-
District E	-	10
District F	-	-
District G	465	72
District H	434	13
TOTAL	2,231	288

Figure 1. Histogram of student postscores in two district/courses in 2012-13.

Panel A. District D's British Literature Posttest



Panel B. District F's British Literature Posttest

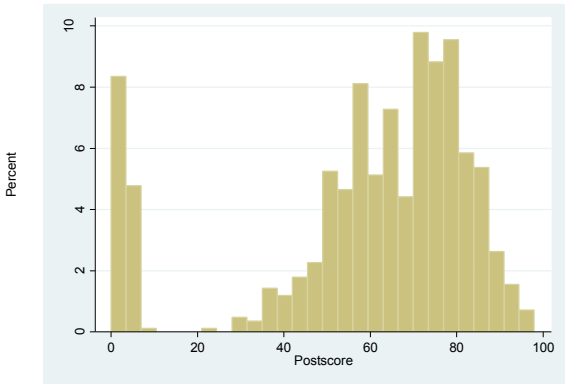
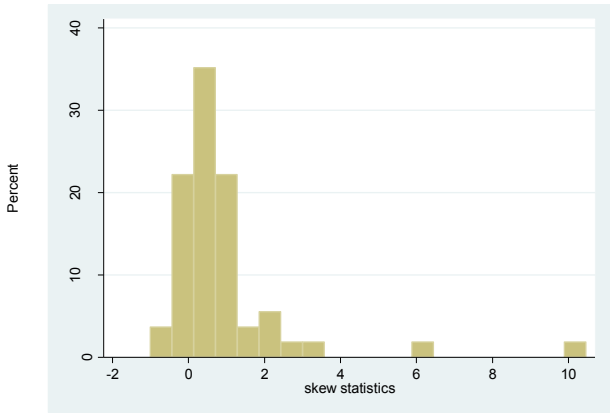


Figure 2. Histogram of skew statistics from 2012-13 assessments (n= 54 district/courses).

Panel A. Pretest skew statistics



Panel B. Posttest skew statistics

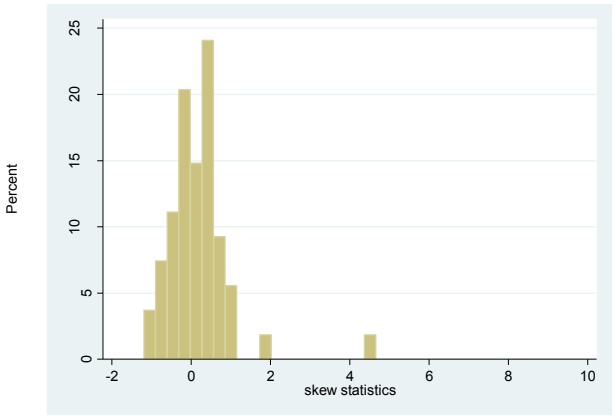


Figure 3. Histogram of student prescores in District H's World History 2012-13 assessment (n=1,477).

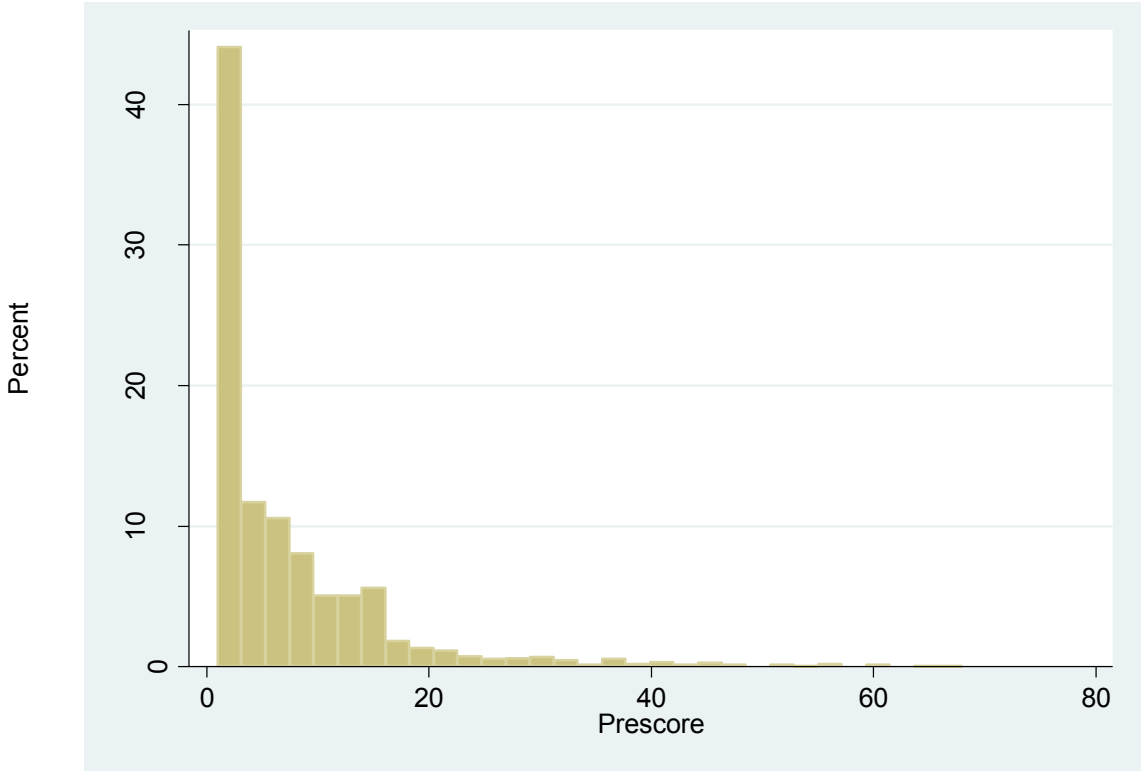


Figure 4. Histogram of prescore-postscore correlation coefficients for 2012-13 assessments (n= 54 district/courses).

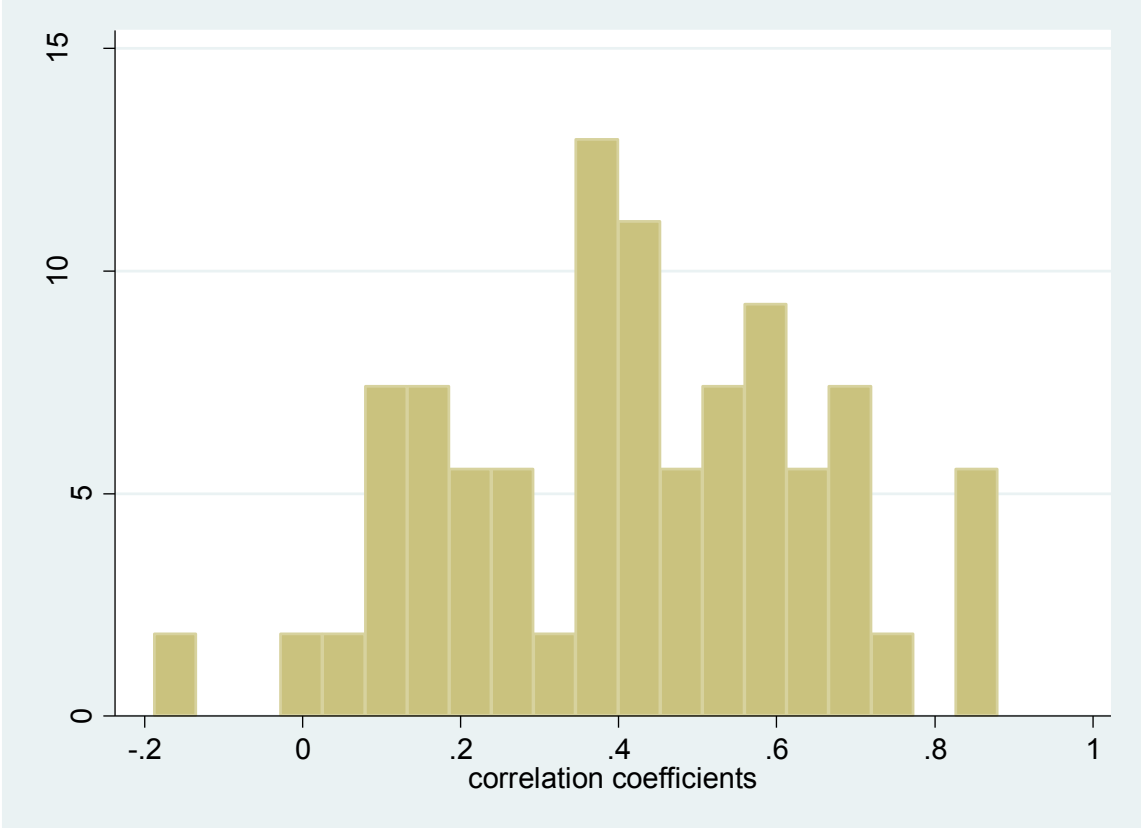
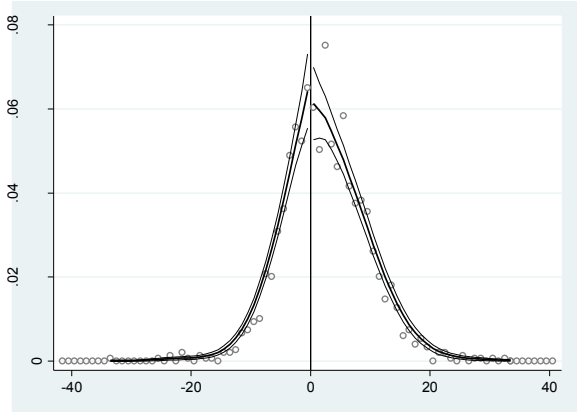


Figure 5a. Density of *post_target* scores in District A's third grade Mathematics course in 2012-13 (Panel A) and 2013-14 (Panel B) with regression lines based on local linear smoothing and 95% confidence intervals.

Panel A. 2012-13 scores



Panel B 2013-14 scores

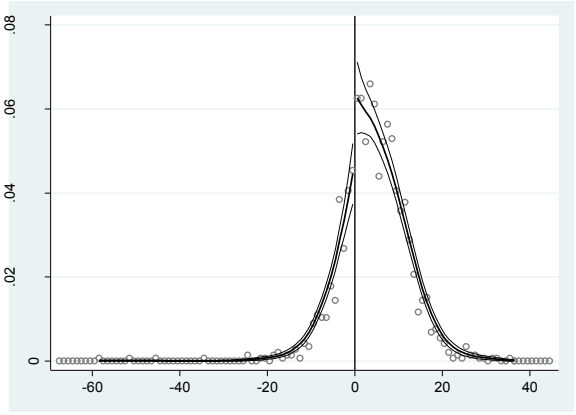
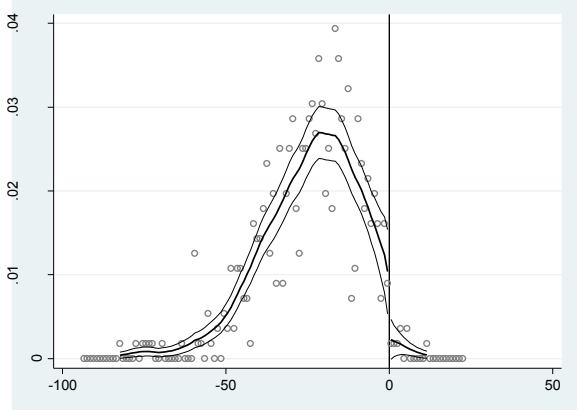


Figure 5b. Density of *post_target* scores in District H's Literature and Composition course in 2012-13 (Panel A) and 2013-14 (Panel B) with regression lines based on local linear smoothing and 95% confidence intervals.

Panel A. 2012-13 scores



Panel B. 2013-14 scores

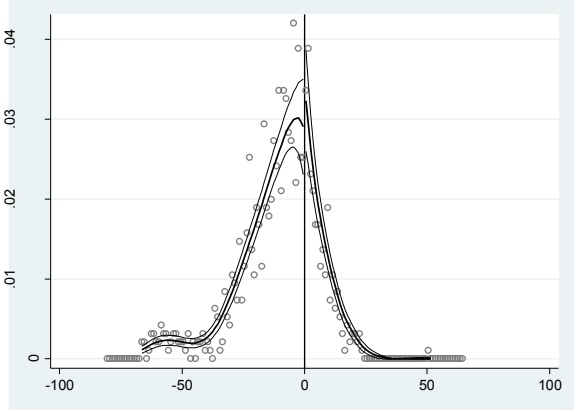
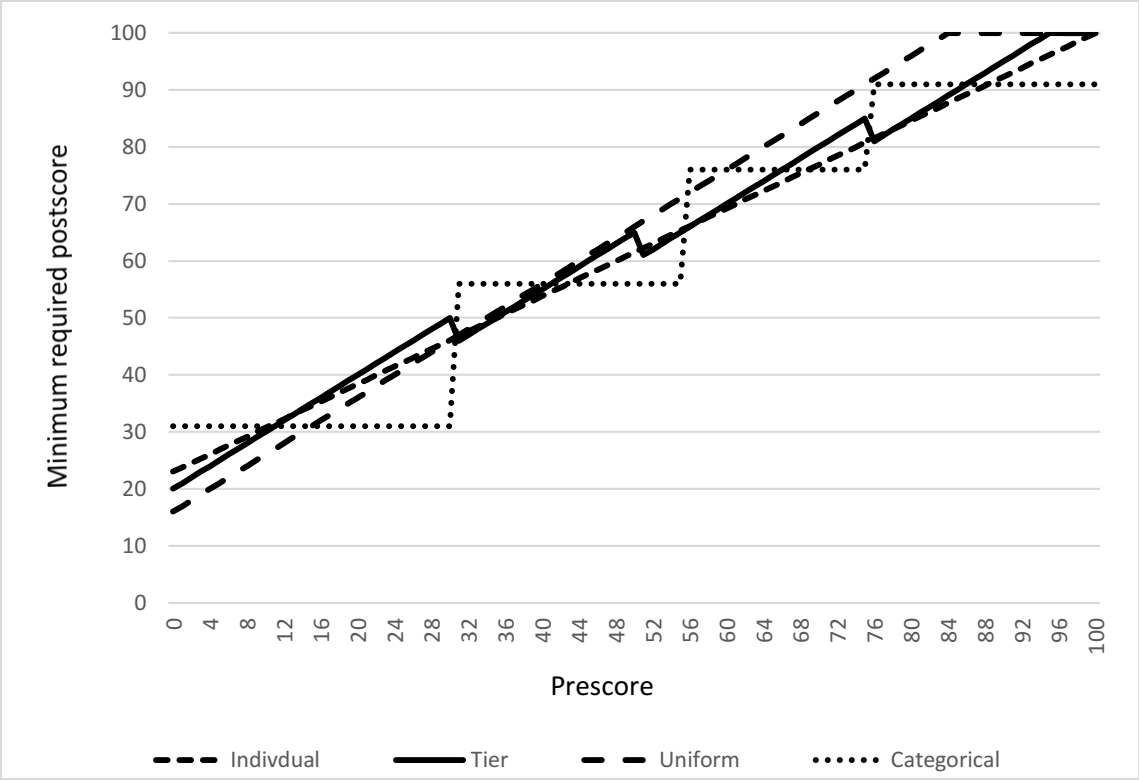


Figure 6. Minimum required postscore as a function of prescore, by target type.



Note. Lines for each target type have been graphed based on target formulations presented in Table 3.

Chapter 2: The Comparability, Reliability and Validity of Teacher Scores Based on Student Learning Objectives

Introduction

In Chapter 2, I examine the validity of teacher-level SLO scores in one state. Given the way in which teacher SLO scores are created, there are two primary aspects that require greater investigation. First, because there are different mechanisms for constructing growth targets across classrooms/courses/districts, an understanding of whether SLO scores are providing comparable measures of teacher performance across the state is needed. In fact, comparability of teacher scores is a primary concern among states implementing SLOs (Lachlan-Hache, et al, 2013). Second, because the assessments used to gauge student progress typically vary across classrooms, we might ask whether and how differences in the quality of data that results from different assessments affects evidence on the validity of scores. I discuss each issue in greater detail below.

Sensitivity of teacher SLO scores to student growth targets

The first focus of this chapter is on the sensitivity of teacher SLO scores to differences in student growth target construction. Fundamentally, this investigation is about the comparability of teacher SLO scores within a state, a key element in ensuring scores are valid for use in an accountability system (AERA, APA, NCME, 2014). The issue of comparability of scores in an SLO framework is quite nuanced (Lachlan-Hache et al., 2013; Marion and Buckley, in press). On the one hand, a fair system requires that a teacher's score is not contingent upon the classroom/school/district in which she/he happens to teach. On the other hand, comparability requires some degree of

standardization, which can create misalignment between the SLO components and the classroom curriculum/student body and interfere with the instructional goals of the SLO. As a first step towards framing this issue, a descriptive look at the extent to which alternative student growth targets classify teacher performance differently, as well as a better understanding of the extent to which teacher scores vary due to the standards imposed by districts through student growth targets, is necessary. I describe each aspect in greater detail below.

Alternative target types. An unanswered question is whether teacher SLO scores in one state are sensitive to different methods for constructing growth targets. While one district may require every student to gain the same number of points from pre- to post-test, another may require every student to reduce the distance between his/her prescore and the maximum number of points on the test by a certain percentage. Alternative growth target formulations can lead to different conclusions about student performance, as demonstrated in Chapter 1 along with prior research (Castellano and Ho, 2013; Goldschmidt, Choi & Beaudoin, 2012; Hoffer et al., 2012), and therefore, it is reasonable to expect that alternative formations will also impact the comparability of teacher scores. In a similar line of research, economists have examined whether alternative value-added models alter the rank of teacher performance (Goldhaber and Theobald, 2013; Wright, 2010; Lockwood et al., 2007; McCaffrey et al., 2004). While strong correlations were generally found across alternative VAM models, a substantial number of teachers, particularly at the tails of the distribution, were affected by model choice. Despite such evidence about the impact of model choices on current school and teacher accountability metrics, no research to date has examined how inferences about teacher SLO

performance are impacted by the method used to set the amount of student growth required of students.

Comparable standards. While variation in SLO components is to be expected, as SLOs are tailored to a particular teacher's curriculum and classroom makeup, variation in the standard imposed by growth targets can mean that a student who meets his/her SLO in one classroom may not do so in another classroom, even if she/he learns the same amount and is administered the same assessment (Lachlan-Hache et al., 2013). This has implications for teacher scores, given that they are calculated by aggregating student SLO attainment within a course. In particular, the extent to which the standard districts impose through student growth targets affects the comparability of teacher scores across districfts and the degree to which district changes to the standard over time impacts inferences of teacher performance is unclear To examine differences in the standards set by districts and disentangle actual changes in teacher scores over time from district changes to the student growth target, an examination of how teacher scores compare under district-set targets and when the same target is applied uniformly to the underlying data is required.

Reliability and validity of teacher SLO scores

The second focus of this chapter is on the reliability and validity of teacher SLO scores and the extent to which assessment data quality biases those results. A comprehensive validity argument approach, in line with Kane (2006), is beyond the scope of this report, due to both data and time restrictions. However, if teachers' SLO scores are reflective of their underlying ability to help students learn, and assuming that teacher effectiveness is relatively stable over time and across subject areas (Loeb and Candelaria;

2013; Goldhaber and Hanson, 2012), SLO scores should be fairly stable over time and across courses. Additionally, SLO scores should converge with other metrics that purport to measure a teacher's effectiveness based on student growth and diverge from measures of classroom demographics. The final two aspects have important policy implications as well: if, for a single teacher, two metrics of performance based on student growth point in different directions, it would be a difficult endeavor to produce a final rating for that teacher. Moreover, if teachers who work in classrooms with lower-performing student subgroups consistently have lower SLO scores, it may disincentivize effective teachers from working in schools serving these populations. I expand on each of these issues below.

Reliability of SLO scores. Any metric used in a high-stakes teacher evaluation system, such as SLOs, should provide a consistent measure for teachers over time and across courses (Bill & Melinda Gates Foundation, 2010). Investigations into the reliability of SLO scores are conceptually similar to a line of research on VAM-based metrics, which also provides a context for the types of reliability we might expect in the current investigation. Within-teacher *across-year* correlations of teacher value-added scores, a form of test-retest reliability, generally range from 0.5 and 0.7 ((Loeb and Candelaria, 2013; McCaffrey et al., 2009). Within-teacher *across-course* correlations of value-added scores, a form of alternate form reliability, generally range from 0.2 to 0.4 (Loeb and Candelaria, 2013; Bill & Melinda Gates Foundation, 2010). These findings suggest that the inclusion of different students in the analysis from year to year and course to course, along with differences in teacher expertise (both over time and across courses taught) may affect the consistency of teacher VAM scores. Since estimates of the

stability of SLO scores will also be influenced by factors involving variation in student body and teacher expertise, the estimates found with value-added scores provide an upper bound on the estimates I would expect to find with teacher SLO scores.

In addition to variation in student body and teacher expertise, the use of different assessments can bias correlational estimates measuring the stability of VAM metrics (Lockwood et al., 2007; Sass, 2008). Papay (2011) found within-teacher within-course stability from VAM scores based on alternative assessments to be as low as 0.15 in certain cases, which he attributed almost entirely to measurement error inherent in the alternative assessments. Likewise, the MET project, which examined the relationship between within-teacher value-added scores, based on different classrooms and different assessments (with different stakes attached to each assessment), found correlations between 0.18 and 0.38 (Bill & Melinda Gates Foundation, 2010). This research is particularly relevant to that of SLOs because the assessments used for SLOs are quite new, and in general, have not been validated for use in a teacher evaluation system (Herman, Heritage and Goldschmidt, 2011; Steele et al., 2011). In fact, in Chapter 1, I found evidence of anomalous test score distributions and a lack of association between student pre and post-scores on SLO assessments, issues which call into question the validity of student growth scores. As such, any investigation into the reliability of teacher SLO scores must also take into account the extent to which correlations are biased by low-quality assessments.

Convergent validity of SLO scores. In addition to providing reliable estimates, teacher SLO scores should converge with alternative metrics purporting to measure a teacher's contribution to student growth, as a matter of validating the measure and as a

matter of policy. One such metric is a mean Student Growth Percentile (MGP) score, which is being used in place of value-added scores in at least seven state evaluation systems, including the state whose SLO data I am examining (D. Betebenner, personal communication, December 29, 2014). Strong within-teacher across-course correlations between the two metrics would provide evidence that SLOs are measuring a teacher's contribution to student growth in the same way as alternative metrics based on standardized assessments.

However, like with the stability of teacher SLO scores, the within-teacher relationship between SLO scores and alternative metrics of performance from different courses may be confounded by within-teacher differences in content knowledge and pedagogical techniques across courses, as well as classroom peer effects (see Loeb and Candelaria, 2013; Bill and Melinda Gates Foundation, 2010). The relationship between SLOs and alternative metrics of performance may be further confounded with differences in the metrics used to derive each measure, as well as differences in the assessments used to calculate each score, particularly since newly created assessments for SLOs may not have technical quality comparable to that of standardized assessments (Steele et al., 2011; Herman et al., 2011). In REACH's pay-for-performance system, for example, researchers found correlations between the percentage of teachers who met their SLO targets and school wide performance results based on standardized assessments ranging from -0.09 in ELA to 0.50 in Science (Schmitt et al., 2014), suggesting that differences in the assessments used across subjects may be driving the mixed results. Given these confounding factors, it is difficult to hypothesize about the strength of the relationship

between teacher SLO and MGP scores; however, it will be bounded by the within-teacher across-course stability of both SLO scores and MGP scores (Glazerman et al., 2011).

Influence of classroom makeup on SLO scores. A final question is the extent to which teacher SLO scores are influenced by factors beyond a teacher's contribution to student learning. While there are many such factors, including school culture, student home life, etc., the makeup of the classroom is of particular concern since any system that penalizes teachers with higher percentages of at-risk students will disincentivize effective teachers from working with these populations. As such, a negative relationship between teacher SLO scores and, for example, the percentage of students with disabilities, might suggest that the system is treating certain teachers unfairly.

In general, literature typically finds correlations between VAM estimates and classroom demographics (i.e., percentage of free and reduced lunch students) to be between -0.02 and -0.25; the higher end of this range (-0.25) suggests the presence of statistical bias (Goldhaber and Theobald, 2013; Hill et al., 2011; Ballou et al., 2004; Kupermintz, 2003; McCaffrey, et al., 2004; Tekwe et al., 2004). However, a negative relationship between teacher scores and percentage of at-risk populations does not necessarily indicate statistical bias if highly effective teachers tend to sort into more advantaged schools (Hill et al., 2011; McCaffrey et al., 2003).

Research Questions

Data from one's state SLO system, in which the assessments and growth targets vary across course and districts, are used to address the issues raised above. When appropriate, I compare results from 2012-13 to 2013-14, in order to examine whether the

quality of scores improves in an intended high-stakes year and as districts develop their systems for conceptualizing and implementing SLOs. I ask:

1. *How sensitive are teacher scores to different methods for constructing student growth targets? Specifically:*
    a. *How do alternative target types applied to the underlying data affect the relative position of teacher scores?*
    b. *How do district-level SLO scores change when student growth targets are applied uniformly to every course in the sample in each year, and what does this say about variation in the standard applied across districts and within districts over time?*
2. *What does the evidence suggest regarding the reliability and validity of teacher scores, and to what extent is this evidence biased by the quality of student assessment data? Specifically:*
    a. *What is the within-teacher across-year stability of SLO scores (test-retest reliability) and what is the within-teacher within-year across-course reliability of teacher SLO scores (alternate forms reliability)?*
    b. *What is the convergent validity of teacher SLO scores based on the relationship between teacher SLO scores and teacher MGP scores for teachers who have both metrics derived from different courses?*
    c. *To what extent are SLO scores in a high stakes year correlated with classroom make-up, including average prescore, percentage of students with disabilities (SWD) and percentage of English Language Learners (ELL)?*

Data and Method

Data

Two state-produced datasets are used to address each research question. The first contains student-level SLO prescores, postscores, target scores and demographic data from the 2012-13 and 2013-14 school year, as described in Chapter 1.[32] The second dataset contains teacher-level MGP scores from 2012-13 and 2013-14.

To address the research questions proposed in this chapter, I transformed the student-level dataset into a teacher/course-level dataset containing teacher/course SLO

[32] While I do not have information on the target formation used for courses in each year, I was able to determine this by examining the relationship between target scores and prescores for every student within a district/course, given that student target scores within a district/course are calculated as a function of student prescores for each target formulation.

scores and classroom demographics, whereby courses are nested within teachers. Using teacher and course IDs, I calculated the percentage of students with disabilities (SWD), the percentage of English Language Learners (ELLs), and the average standardized prescore for each teacher/course.

Each teacher's SLO score for each course was calculated as the percentage of each teacher's students whose postscore met or exceeded their target score. Prior to calculating teacher SLO scores, I excluded students if they had (a) missing prescores, since their target score could not be verified, (b) postscores greater than the maximum number of points on the test, which would suggest an error in grading, or (c) missing postscores, since postscores were needed to determine each student's SLO designation.[33] See Appendix C, Table C1 and C2, for a summary of excluded student data. Further, I used only corrected target scores, as determined in Chapter 1, to calculate student SLO attainment.[34] Finally, I excluded teacher SLO scores from the dataset if the district/course contained less than 200 students with SLO scores, given the issues that can arise with small sample sizes.[35]

After creating the teacher/course SLO dataset, I merged MGP scores for every teacher who had an MGP score from a matched course, whereby I considered a matched

---

[33] A special case existed with District E, where in several subjects, the state data did not match the district data. Where discrepancies existed, I relied on district data instead of state data.

[34] In Chapter 1, I found cases of target misclassification, whereby students within a district/course who had the same prescore and thus should have the same target score did not (1.9% of students in 2012-13 and 0.2% of students in 2013-14 had misclassified target scores). In these instances, I determined the correct target classification that was applied to the majority of students in the district/course and applied it to all students with incorrect target scores in order to produce corrected target scores.

[35] Because the courses selected for the sample were core courses, each course generally contained hundreds if not thousands of students with SLO scores. Therefore, when a course contained scores for less than 200 students, it suggests that either (1) the data wasn't fully reported for that course or (2) the SLO assessment was not administered uniformly for all classrooms covering the course. For example, in District E's British Literature course, 133 students had scores from 2012-13 but 1,791 students had scores from 2013-14. This increase was likely due to the fact that the assessment was not administered or reported uniformly in the pilot year of implementation.

course to be that which covered the same subject as the course from which a teacher's SLO score was derived. Therefore, teachers with matched SLO scores and MGP scores in my dataset are teachers with scores from different metrics based on different courses (with generally different groups of students) but covering the same subject. A list of the matched SLO and MPG courses is provided in Table 11. Teachers with both scores span all primary subjects (Mathematics, ELA, Social Science and Science) at both the elementary school level and high school level.[36]

[insert Table 11 here]

Sample

In the 2012-13 sample, there were a total of 3,493 teachers with SLO scores, and in the 2013-14 sample, there were a total of 4,588 teachers with SLO scores. These frequencies include teachers who may be counted more than once if they teach multiple courses with an SLO.[37] Table 12 displays frequencies of teachers with SLO scores in each district/course in the sample in each year. In general, the number of teachers with SLO scores increased from 2012-13 to 2013-14 as districts ramped up their SLO implementation.[38] This is particularly true of District E, which delayed full implementation of SLOs (i.e., test administration and reporting) until 2013-14. On the other hand, District F has a smaller number of teachers with SLO scores in 2013-14

[36] Note that no teacher has both metrics at the middle school level in the sample due to the fact that standardized assessments are administered in all core academic courses at this level.
[37] Twenty-six percent of teachers from 2012-13 have multiple SLO scores, and 21% of teachers from 2013-14 have multiple SLO scores.
[38] For example, some districts chose to administer an SLO or report scores in 2012-13 for only those teachers who did not have an MGP score. Alternatively, some districts were not ready to implement an SLO in every non-tested course in 2012-13, due to issues with assessment development/selection, and were given leeway by the state.

compared to 2012-13, since this district only submitted to the state student scores for a maximum of two courses per teacher.

[insert Table 12 here]

In 2012-13, of the 3,493 teacher/courses included in the sample, 25% had SLO scores equal to zero, where a score of 0 means that 0% of a teacher's students met their SLO target in the given course. The remainder of the distribution of teacher-level SLO scores appears uniformly distributed, with fairly equal percentages of teachers receiving scores across the SLO score distribution (see Figure 7, panel A). The average teacher-level score is 35.9 (i.e., on average, 35.9 percent of each teacher's students met their target within a district/course), with a standard deviation nearly as large (34.6).

The distribution of the 4,588 teacher-level scores in 2013-14 looks quite similar to 2012-13 (see Figure 7, panel B); however in year two, fewer teachers received a score of zero (15.0% compared to 25%) and a greater number of teachers received a score of 100 (7.2% compared to 4.6%). The average teacher-level score in 2013-14 is higher than in 2012-13, at 46.8, with a standard deviation of 33.4.

[insert Figure 7 here]

Data analysis

Sensitivity of teacher SLO scores to student growth targets. Different districts use different target formulations with varying rigor, and thus teacher scores may not be comparable across the state. Therefore, this section has two components: (1) a comparison of the relative ranking of teacher scores under alternative target types; and (2) a comparison of average teacher scores within each district under district-set targets and under an alternative target applied uniformly to the data in both years. To address

both, I applied the four alternative target formulations described in Chapter 1 to the data (see Table 3 in Chapter 1 for target formulations). Recall that each target type was calculated such that, when applied to the underlying 2012-13 data, 52% of students in the sample met each alternative target. This holds constant the level of rigor (i.e., the standard) across alternative targets types.[39]

To examine whether alternative target types altered the relative position of teacher scores, I calculated four new sets of teacher-level scores. With these four sets of alternative teacher scores, I compared the relative ranking of 2012-13 teacher scores under each alternative target type using Spearman rank correlations. Because the target rigor was held constant, these analyses allow for an examination of how teacher scores change solely as a function of target type.

To examine how the rigor of growth targets vary across districts and within districts over time, I directly compared district-level average teacher scores under the district-set target and under one alternative individual target requiring 23% growth (i.e. each student's target score is based on closing the distance between his/her prescore and the maximum number of points on the test by 23%).[40] Importantly, applying the alternative individual target to the underlying data ensures that the type *and* rigor of the student growth target is held constant within and across districts. This allowed for an

[39] Note this percentage was based on applying each target formulation to every student in the sample of district/courses, and did not exclude students from courses with less than 200 students.

[40] One cannot objectively compare differences in the rigor applied by districts by simply comparing the percent of students meeting their target across districts, since this percentage will also be a function of actual student ability, teacher effectiveness, and the difficulty of the assessments. Instead, applying the alternative individual target serves as a barometer with which to compare the standard set by districts across districts. While any alternative target would have worked for this analysis, I chose the individual target formulation because (1) the indidvual target was the most commonly used by districts; and (2) the individual target formulation allows for the most straightforward conception of rigor, which is a direct function of the percentage growth required.

examination of: (a) the extent to which districts vary in the standard they hold students to, by comparing across districts the shift in teacher-level scores from the district-set target to the alternative individual target in a given year; and (b) the extent to which district changes to the student growth target from year one to year two impacted the change in teacher scores, by comparing the change in teacher scores over time under district-set targets to the change in teacher scores over time under the alternative individual target.

Reliability and validity of teacher SLO scores. I investigated the reliability of SLO scores by examining (a) the within-teacher across-year SLO correlation for teachers who have SLO scores from the same course in 2012-13 and 2013-14, based on test-retest reliability; and (b) the within-teacher across-course SLO correlation for teachers who teach multiple courses in a given year where an SLO was administered, based on alternate forms reliability.

To examine the convergent validity of teacher-level scores, I calculated the Pearson correlation between teacher SLO scores and MGP scores (the SLO-MGP correlation) for teachers who have both scores from matched courses (discussed above). Doing so removed subject-specific teacher-expertise as a confounding factor since a teacher's MGP and SLO score will come from courses that although different, cover the same subject.

To investigate whether teacher SLO scores are influenced by classroom makeup, I examined the relationship between a teacher's SLO score and the average classroom standardized prescore, the percentage of students within each classroom designated as ELL, and the percentage of students within each classroom designated as SWD.

In Chapter 1, I investigated the suitability of assessments in the sample by looking for evidence of (a) anomalous test scores distributions; (b) skew or spikes in the data; and (c) low pre-postscore correlations. In total, 50% of district-course combinations in 2012-13 and 32% district-course combinations in 2013-14 contained data quality issues. Based on results from this investigation, 60% of teacher scores from 2012-13 and 28% of teacher scores from 2013-14 came from district/course combinations with poor data quality.[41] In order to examine the degree to which the quality of the assessments affects the reliability and validity of SLO scores in this chapter, I investigated how correlations described above shifted when the sample of teachers was limited to only those in district/courses containing assessments with acceptable statistical properties (i.e., the reduced sample of teachers).

Results

Sensitivity of teacher SLO scores to student growth targets

To begin the analysis of the sensitivity of teacher scores to target formulation, I first describe trends in teacher performance in each district based on district-set targets. As illustrated in Table 13, there is a great deal of variation in average teacher scores across districts within each year, with performance varying across districts by as much as 53 percentage points. Based on a variance decomposition, a majority of variation in teacher-level scores in 2012-13 occurs within district; however 31% occurs between districts, a relatively large amount.[42] From 2012-13 to 2013-14, the between-district

[41] Note that in both the 2012-13 and 2013-14 sample, average teacher performance was lower in the full sample of teachers compared to the sample of teachers with adequate student assessments (i.e., the reduced sample) (35.9 and 45.6, respectively in 2012-13 and 46.8 and 49.8, respectively in 2013-14), suggesting that poor data quality created a downward bias in teacher scores, particularly in year one.

[42] I fit an unconditional random intercepts multi-level model of the following form: $SLO_{td} = \alpha + \varepsilon_{td} + \delta_d$. Using maximum likelihood estimation, the SLO score for each teacher *t* in each district *d* provides a fitted estimate α of the mean teacher SLO score in the sample. Two error terms are specified; a teacher-level

variation decreased from 31% to 26%, suggesting slightly greater comparability over time in how districts were creating SLOs.

Moreover, on average, teachers scores were higher in year two compared to year one. In fact, of the 1,568 teachers with scores in the same course in both years, 74% of teachers had an SLO score in 2013-14 that was equal to or higher than his/her score in 2012-13.[43]

[insert table 13 here]

The questions arising from the differences in teacher performance across districts and over time lead naturally to an inquiry about the effect of district-set targets on teacher scores. In what follows, I address two specific questions: (1) How do alternative target types applied to the underlying data affect the relative position of teacher scores? (2) How do district-level SLO scores change when student growth targets are applied uniformly to every course in the sample in both years, and what does this say about variation in the standard applied across districts and within districts over time?

How do alternative target types applied to the underlying data affect the relative position of teacher scores? To examine the effect of different target types on teacher performance, I compare average teacher performance in each district based on the

---

error term, ε_{td}, and a district-level error term, δ_d. The variation at the district-level is calculated as the between-district variation squared, divided by the total amount of variation squared: $\frac{{\delta_d}^2}{{\delta_d}^2+{\varepsilon_{td}}^2}$.

[43] The exception to this occurs in District B and District E, where average performance based on district-set targets declined from year one to year. Both of these districts increased the number of courses with an SLO from year one to year two, which suggests that the sample of students/teachers in each district in 2012-13 was not fully representative of teacher performance in 2012-13. For example, in year one, only two courses in District E had sufficient student n-sizes to create teacher scores to include in the sample: high school Literature and Comprehension and grade three Mathematics, By year two, teacher scores from eight courses had sufficient n-sizes to include in the sample. Therefore, the number of teachers included in the sample increased from 46 in year one to 673 in year two. Likewise, teachers from five courses in District B were included in the 2012-13 sample, while teachers from seven courses were included in the 2013-14 sample.

application of the four alternative targets to the 2012-13 underlying student data. As displayed in Table 14, average scores diverge across target types within districts by as much as 7 points, with the greatest amount of divergence typically occurring between scores based on a categorical target and the three remaining target types. However, basing scores on a categorical target will not consistently rank teachers higher or lower, as evident by the fact that the average difference between the percentage of students meeting their target under the individual target and under the categorical target, across all districts – whereby each district is weighted equally – is zero.

[insert Table 14]

Table 15 shows Spearman Rank correlations based on teacher scores under each alternative target formulation. While overall, correlations are fairly high across alternative target types, indicating that teacher rankings will remaining mostly the same under each alternative target, any correlation less than one suggests that some teachers will be ranked differently if a different target type were to be applied. As expected based on student results presented in Chapter 1, correlations tend to be the lowest between teacher scores based on the categorical target and the remaining three target types.

[insert Table 15]

To illustrate the effect of alternative target types on the relative position of teacher scores, Table 16 provides the percentage of teachers who would fall into each quartile of the score distribution under an individual student growth target and a categorical student growth target. The majority of teachers (75%) would be placed in the same quartile under both sets of growth targets. However, a quarter of all teachers would be ranked differentially under one set of growth targets compared to the other. Most of the

differences in quartiles based on alternative student targets are one-off (i.e., a teacher would fall in the 2nd quartile under one growth target and the 3rd quartile under the other) while 2% of teachers would rank quite high under one growth target but quite low under the other.

[insert Table 16]

Results from Tables 14-16 suggest that applying alternative targets with the same level of rigor to student data will non-uniformly shift teacher scores for a small percentage of teachers. Shifts will be greatest when switching from an individual, tiered or uniform target to a categorical target. That being said, there is no particular reason to expect that the categorical target will consistently rank teachers lower or higher, on average. Whether a teacher looks better or worse under a particular target is a function of where in the distribution a teacher's students' test scores fall.

How do district-level SLO scores change when student growth targets are applied uniformly to every course in the sample in both years, and what does this say about variation in the standard applied across districts and within districts over time? Recall that districts were instructed by the state to choose both the type and rigor of the target for a given course, which could vary from course to course. As displayed in Table 17, individual student growth targets were the primary target type required by districts in 2012-13 and 2013-14. Districts tended to set similar target formulations for each course within the district and primarily kept the target type used for each course the same from year one to year two. However, there is still quite a bit of variation in the type and rigor of growth targets applied within and across districts.

[insert Table 17 here]

I compare average teacher scores within each district based on district-set targets and the individual alternative target, which requires students to reduce the gap between their prescores and the maximum number of points on the posttest by 23%. Importantly, since applying the alternative individual target to the underlying data holds constant the standard required of districts, it provides a barometer with which to judge the standard imposed by the district-set targets. As such, districts with higher average teacher scores under the district-set target compared to the alternative individual target set, on average, an easier standard relative to districts with lower average teacher scores under the district-set target compared to the alternative individual target.

Table 18 displays average teacher scores under the district-set target and the alternative individual target for each district in each year. In both years, in Districts A and G, the average percentage of teachers' students meeting their target is *higher* under the district-set targets compared to the alternative target. In Districts C, E, F, and H the average percentage of teachers' students meeting their target is lower under the district-set targets compared to the alternative target. [44] This means that Districts A and H set an easier standard compared to Districts C, E, F and H. In terms of comparability of teacher scores across districts, this means that among two teachers, one from the first set of districts (i.e., Districts A and G) and one from the second set of districts (i.e., Districts C, E, F, and H), whose students learned the same amount throughout the course of the year, the teacher from the former group, by virtue of being in a district with easier standards, is more likely to be deemed effective by the state's teacher evaluation system in both years.

[44] District D shows a positive shift from the district-set to the alternative target in 2012-13 suggesting relatively harder district-set targets in this year, and a negative shift in 2013-14 suggesting relatively easier district-set targets in this year. Conversely, District C set relatively harder targets in 2012-13 compared to 2013-14. I discuss both of these special cases in further detail below.

[insert Table 18]

The final row of table 18 illustrates the overall extent to which district-set targets were harder or easier than the alternative individual target requiring 23% growth in each year. In 2012-13, the alternative individual target required a less rigorous standard, on average, than the targets set by the district; evident by an upward shift in overall performance under the alternative individual target compared to the district-set target. However, in 2013-14, the alternative individual target was slightly more rigorous than the targets set by the districts, on average, evident by the drop in average performance under the alternative individual target compared to the district-set target. This suggests that district modifications to the growth targets in year two had the primary effect of decreasing the rigor of student growth targets from year one to year two. In fact, in year one, many districts required, for example, an individual target whereby students were expected to close the gap between their prescore and the maximum points possible on the posttest by 50% or even 65%, but in year two, districts generally required a lower percentage growth of between 20% and 33%. Changes that districts made to the student growth targets between year one and year two not only led to a decrease in rigor required of students, but also increased the comparability of standards across districts, as evident by the fact that the correlation between scores based on district-set targets and the scores based on the alternative target is 0 in year one and .5 in year two.

Given that that districts typically changed the rigor (and to a lesser extent the type) of growth target required of students, an unanswered question is how the changes to student growth targets over time affected the change in teacher SLO scores from year one to year two. In order to isolate the effect of changes to the target on changes in teacher scores

over time, I reconfigure Table 18 as Table 19 to explicitly compare the change over time in teacher scores under district-set targets to the change over time in teacher scores under the alternative individual target. In other words, I examine how average teacher performance in each district would look if the type and rigor of the target remained the same over time.

[insert Table 19]

In general, the improvements in district-level average teacher scores over time appear at a lower rate when the alternative individual target is applied to the data in both years.[45] For example, teacher scores in District A improved by 27.31 percentage points under the district-set target, but only 1.23 percentage points under the alternative individual target. Likewise, teacher scores in District G improved by 47.88 percentage points under the district-set target but only 4.67 percentage points under the alternative individual target. This suggests that much of the increase in teacher scores under district-set targets is the result of districts modifying the rigor of the targets from year one to year two.

Why did districts modify their targets from year one to year two? First, state personnel indicated that districts recognized that the growth required of students based on their original targets was untenable due to the large number of students failing to meet the target in year one. While the state required that student growth targets be established in an objective manner, based on prior years' data, it is difficult to a priori determine how much students are capable of growing on these (mostly) newly developed assessments (Lachlan-Hache et al., 2012). In District H, for example, the average teacher score under

[45] General trends under the alternative target remain similar when I limit the sample to only those teachers who taught the same course in both years (n=1,593).

the district-set target in 2012-13 was only 14.82 (i.e., on average, 14.82% of teachers' students met their targets) but in 2013-14, it was 62.47. As such, District H appears to have decreased the rigor of the target in year two given the low number of students capable of meeting their target score in year one.

The difficulty that districts may have had in understanding how the standard imposed by the target would affect teacher scores is illustrated by District E. This district introduced two new SLO courses in year two, with individual targets requiring 50% growth. This was a harder student target than (a) what most districts were using by year two (i.e., an individual target requiring the gap in student performance to close by 20-33% growth); and (b) the alternative individual target I imposed (23% growth). In fact, the introduction of these new targets in District E created the perception of a decline in teacher performance over time, which reversed under the alternative individual target that I applied. Because of this, I suspect that the district will decrease the level of rigor of the target for these two courses in the 2014-15 school year.

Second, it appears that changes to the assessments that districts made from year one to year two necessitated modifications to the rigor of the target. For example, District D changed the assessments used for third grade Mathematics and ELA from 100-point locally-developed assessments to 1400-point commercially-developed assessments in 2013-14. On the third grade ELA assessment, the average standardized gain in 2012-13 was 2.2 but under the new assessment in 2013-14, the average standardized gain was only 0.30. The district also modified the student growth target from an individualized target requiring 50% growth in 2012-13 to a uniform target requiring a gain of 75 points from pre to post-test in 2013-14. Without this adjustment, aggregate teacher performance in

this district would have declined substantially over time (evident by the decline over time in teacher performance under the alternative individual target which does not account for the difficulty of these new assessments).

These results suggest that district modifications to their student growth target formulations, possibly to adjust for the difficulty of the standard imposed by the target or the difficulty of the assessment, had a substantial effect on changes in teacher's scores over time. That being said, teacher changes in scores under the alternative individual target, while keeping the target type and rigor constant over time, should not necessarily be taken as the "true" change in teacher scores over time. This is because the change in teacher scores under the alternative individual target is still confounded with both changes in the composition of the student body, as well as changes in the assessments administered. District D provides a telling example: had they not altered their targets in third grade ELA and third grade Mathematics in 2013-14 to account for the change in the difficulty of the assessments, teacher scores would have been substantially lower in 2013-14 compared to 2012-13.[46]

Reliability and validity of teacher SLO scores

Inter-temporal stability. To investigate the reliability of teacher scores, I first report evidence on the inter-temporal stability of scores for teachers who have an SLO score from the same course in both years (n=1,593). SLOs appear to provide a moderately stable measure of a teacher's performance over time, with an across-year correlation of 0.46 (p <.001). However, given what is known about the quality of assessment data in

[46] In 2013-14 in District D, 49.5% and 53.1% of students met the district-set growth targets in third grade ELA and third grade Mathematics, respectively. If an individual target requiring 23% growth was applied (which is a less rigorous growth target than the district-set individual growth percentage of 50% for these courses in 2012-13) only 6.8% and 8.5% of students would have met their growth targets.

this state, as discussed in Chapter 1, as well as changes over time to the student targets described in the prior analysis, it is important to disentangle the influence of each of these factors on the reliability of teacher scores over time.

When I limit the sample to teachers in courses with adequate assessment quality in each year (n=595), the correlation estimate drops to 0.18 ($p<.001$). I posit that the primary driver for the lower across-year correlation in the reduced sample is the reduction in bias due to low quality assessments which likely inflated the correlation coefficient in the full sample. In particular, the high correlation in the original sample appears to be "zero-inflated", due to the large number of teacher SLO scores equal to zero in both years. Figure 8 illustrates this clearly by showing a scatter plot of within-teacher SLO scores from the 2013-14 school year (y-axis) and 2012-13 school year (x-axis) in the full sample (Panel A) and reduced sample (Panel B). There are a greater number of teachers with low scores (and high scores) in each year in the full sample compared to the reduced sample, which appears to have biased the correlation in the full sample upwards. Removing the scores of teachers with data based on poor assessment data tends to remove many of these anomalous scores

[insert Figure 8 here]

Given the finding discussed previously in this chapter, regarding district changes to both the assessments and targets from year one to year two, it is likely that part of the low inter-temporal stability found among teachers in the reduced sample is driven by changes to teacher rankings due to district changes to the targets and assessments. While I cannot control for changing assessments, I can estimate the across-year correlation of teacher scores when every student score is derived from the same target formulation: an

individual target requiring 23% student growth from prescore to maximum number of points on the test. The across-year correlation, in the reduced sample of teachers with scores from adequate assessments, based on the same target formulation applied uniformly to the data, is 0.48 p<.001 (n=595).[47]

These findings suggest that the both the quality of assessments and differences in the targets applied within courses across years substantially affect the inter-temporal stability of SLO scores. However, it not entirely obvious which way the original estimate will be biased, since the across-year correlation in the original sample will be a function of how poor quality assessments bias scores within and across years, as well as the extent to which the changes to course targets from one year to the next rank teachers differently.

Across-course reliability. To investigate the within-teacher across-course SLO reliability, I limit the sample to teachers who have multiple course-specific SLO scores derived from courses with acceptable assessments, given that poor assessment data quality appears to bias the across-year stability correlation. The 2012-13 across-course correlation for teachers (n=257) is 0.58 (p <.001). In 2013-14, the across-course correlation for teachers (n=561) is higher, at 0.65 (p <.001).[48] These results indicate that teacher scores are moderately stable across courses, and suggest that districts are creating SLOs in year two that hold teachers to a more consistent standard across courses compared to year one.

[47] The inter-temporal reliability estimates are similar when the alternative uniform target, alternative tiered target, and alternative categorical target are each applied to the data.

[48] The 2013-14 across-course correlation is higher for the reduced sample compared to the full sample (0.65 compared to 0.55), despite restriction of range. In 2012-13, the across-course correlation is lower for the reduced sample compared to the full sample (0.58 compared to 0.67).

Convergent validity. To examine the convergent validity of SLO scores in 2012-13, I limit the sample to 794 teachers (23% of the full sample) that have both an SLO score and an MGP score from different courses covering the same subject.[49] Eight-one percent of teachers in the SLO-MGP sample teach at the high school level, and 19% teach at the elementary school level. Forty-eight teachers are counted more than once since they teach multiple non-tested grades and subjects.[50] Teachers in the SLO-MGP sample have an average SLO score of 23.5, over ten points lower than the average SLO score among teachers in the full sample. The average MGP score of teachers in the SLO-MGP sample is 47.3. Finally, while MGP scores are normally distributed, SLO scores among teachers in the SLO-MGP sample are positively skewed, with 38% of teachers scoring a 0 (see Figure 9).

[insert Figure 9]

Within-teacher SLO and MGP scores are weakly yet significantly correlated (0.13, $p<.001$). What does the SLO-MGP correlation mean substantively? When teachers are placed into quartiles based on the ranking of their SLO score and MGP score, only 32% of teachers would fall into the same quartile under both metrics (see Table 20). Importantly, more than 11% would fall in the highest quartile under one metric and the lowest quartile under the other. This means that for over one-tenth of teachers in the

[49] Recall that while approximately 20% of teachers from the state have both metrics, the sample of districts and subjects was selected in order to maximize the number of teachers with both metrics, which is why the sample contains a greater percentage of teachers with both metrics compared to the state.

[50] For teachers who teach multiple SLO courses, the most common cases were 3rd ELA and 3rd grade Mathematics (i.e., a teacher taught both 3rd grade ELA and 3rd grade), British Literature and 10th grade Literature and Composition or World Literature, and American Government and World History. In cases in which teachers teach multiple non-tested courses that vary in subject matter (e.g., third grade ELA and third grade Mathematics), their matched MGP score is different for each SLO course; in cases in which the courses taught cover the same subject (e.g., British Literature and World Literature), their matched MGP scores is the same for each SLO course.

sample, their alternative metrics of student growth are pointing in completely opposite directions.

[Insert Table 20]

As a sensitivity test to ensure that the low SLO-MGP correlation is not due to the differences in the distributions between teacher SLO scores and MGP scores, I created an alternative metric of teacher performance based on an average implied gain. I first calculated a student implied gain for every student included in the teacher sample, whereby the implied gain is the observed student gain in the units of the required gain as specified by the target score: $Implied\ gain = \frac{x_{post} - x_{pre}}{x_{target} - x_{pre}}$. [51] An implied gain of 1 indicates that student's actual growth exactly matches his/her required growth; an implied gain of less than 1 indicates that a student's actual growth falls short of his/her required growth, and implied gain of greater than 1 indicates that a student's actual growth exceeds her/his required growth. I then calculated the average implied gain for each teacher within each course.

Recall that the distribution of teacher-level SLO scores is positively skewed, with a large percentage of teachers receiving a 0. However, as with the distribution of MGP scores, teacher scores based on average student implied gains are normally distributed (see Figure 10). The resulting correlation between teacher-level implied gain averages and MGP scores remains the same as the SLO-MGP correlation, 0.13, suggesting the shape of the SLO distribution and the way the metric is derived does not play a role in the low SLO-MGP correlation.

[51] Note that an implied gain score will be undefined in cases where one's prescore equals one's target score, since the denominator will equal zero. This affects four students in the student-level sample and removes one teacher from the sample (n=793).

[insert Figure 10]

The original SLO-MGP correlation of 0.13 signifies that for every one percentile increase in a teacher's MGP score, his/her SLO score increases by slightly over a quarter of a percentage, on average. However, it is unclear whether this relationship is in fact, linear. In Table 21, I report the MGP-SLO correlation coefficient for teachers grouped according to MGP scores, in order to determine if the MGP-SLO relationship varies as a function of MGP score. Indeed, there appears to be a nonlinear relationship between within-teacher SLO and MGP scores, such that the SLO-MGP correlation among teachers with MGP scores in the lowest third of the MGP distribution is negative but not significant, the SLO-MGP relationship among teachers with MGP scores in the middle of the distribution is small but positive, and the SLO-MGP relationship among teachers with MGP scores in the highest third of the distribution is moderate and positive.

[insert Table 21]

The non-linear SLO-MGP relationship suggests that SLO and MGP scores converge more for teachers at the higher end of the MGP distribution. If we take a teacher's MGP score as a true measure of effectiveness, this suggests that SLOs are not able to accurately distinguish true teacher effectiveness at the lower end of the effectiveness spectrum. While it's not entirely clear why this is the case, part of the reason is likely due to the quality of assessments. In particular, in the 2012-13 school year, lower quality assessments tended to bias teacher SLO scores downward, as evident by the lower average SLO score among teachers in the full sample compared to the sample of teachers with adequate assessments (35.9 compared to 45.5, respectively). Further, average SLO scores are lower among teachers with MGP scores in the bottom third of the distribution

compared to teachers with MGP scores in the upper two thirds of the MGP distribution, as seen in Table 21. As such, teachers in the bottom third of the MGP distribution may be more likely to be in courses where the assessments have poor data quality. Therefore, I next examine the convergent validity of SLO scores when the sample is limited to only those district/courses with adequate assessment data quality (i.e., the reduced sample) in order to remove potential bias from low quality assessments on the relationship between teacher SLO and MGP scores.

When the 2012-13 sample of teachers is limited to teachers from the 27 courses with pre and post-assessments containing adequate statistical properties (n=268), the within-teacher across-course correlation between SLO scores and MGP scores is greater than in the full sample (0.24, p<.001).[52] Notably, the correlation for teachers in subjects with assessments with poor data quality (n=526) is very weak and not significant (0.06, p=0.20). Therefore, it does appear that low quality assessments are biasing the SLO-MGP correlation. This may be due to the fact that in year one, low quality assessments tended to bias teacher scores toward zero. Removing these teachers from the sample, therefore, could improve the overall correlation between SLO scores and MGP scores due to greater variation in the SLO distribution. This is evident in Figure 11, which provides a scatterplot of within-teacher MGP scores (y-axis) and SLO scores (x-axis) for teachers in the full sample (panel A) and teachers in the reduced sample (panel B).

[52] The increased strength in the SLO-MGP correlation in the reduced sample is even more pronounced at the elementary school level. When I disaggregate the data by school-level, the SLO-MGP correlation among teachers with high-quality assessments at the elementary school level is .36 (n=36), p=.03. This strength of this relationship appears to be driven by the reduction of elementary school teachers with low SLO scores from the full to the reduced sample (the average SLO score in the full sample of elementary school teachers is 40 while the average SLO score in the reduced sample of elementary school teachers is 64). The SLO-MGP correlation among teachers with high-quality assessments at the high school level is only .14 (n=232), p=.04) (and the average SLO score among high school teachers increases from the full sample to the reduced sample by only 4 points).

[insert Figure 11]

In 2013-14, 1,063 teachers from the sample have both SLO scores and MGP scores from courses covering the same subject. Sixty-seven percent of teachers in the SLO-MGP sample teach at the high school level, and 33% teach at the elementary school level. Teachers in the SLO-MGP sample have an average SLO score of 36.7, higher than in 2012-13, and an average MGP score of 47.1.

The teacher SLO-MGP correlation in 2013-14 is the same as in 2012-13 (0.13, $p<.001$). Interestingly, reducing the sample to teachers from courses with adequate data quality (n=705) does not improve the correlation as it does with the 2012-13 sample of teachers. Why might this be? First, in 2012-13, removing teachers with low quality assessments tended to remove teachers with scores of zero which biased the SLO-MGP correlation downward. The same is not the case in 2013-14, where average teacher scores among teachers with poor data quality in the sample of teachers with both SLO scores and MGP scores is slightly higher, on average, compared to average teacher scores among teachers with adequate assessment quality. Second, it is possible that the criteria used to determine assessment data quality in Chapter 1 may have been more effective at identifying poor assessments in 2012-13 because the issues were more serious in 2012-13 (e.g., most assessments in 2012-13 contained several indicators of poor data quality; however in 2013-14, assessments identified as having poor data quality generally only contained one indicator – low prescore-postscore correlations) or because there may be other issues occurring in year two, which may distort teacher SLO scores, that the indicators of assessment quality are not fully picking up (i.e., teaching to the test).

Influence of classroom makeup on SLO scores. Finally, I investigated the extent to which teacher-level SLO scores are influenced by the composition of the classroom. For this analysis, I make use of data from the 2013-14 school year only (n=4,359), since 5% of students in the 2012-13 sample are missing ELL and SWD data, which could create misleading classroom percentages.

There is a small but positive correlation between a teacher's SLO score and the average standardized prescore of the classroom, 0.17 (p <.001). The 2013-14 correlation remains similar when the sample is limited to teachers in courses with adequate assessment quality (0.22), suggesting that poor assessment data quality is not driving this finding. Substantively, this relationship between teacher SLO score and average student prescore means that for every 1-unit increase in course average standardized prescores (which range from -3 to 5), teachers, on average, have an SLO score that is 8-10 percentage points higher. This suggests that teachers in classrooms with lower student baseline performance tend to have lower SLO scores, relative to teachers in classrooms with higher student baseline performance.

The relationship between teacher SLO scores and baseline student performance is a somewhat unexpected finding, since student targets are intended to implicitly account for student performance at baseline, by setting growth targets as a function of a student's prescore. Further, students in the sample with lower performance on the pretest grew more than students with higher performance: students in the first quartile of the SLO standardized prescore distribution in 2013-14 in the sample gained, on average, 41 points from pre to post-test, while students in the fourth quartile gained, on average, 11 points. Together, these findings suggest the while it is appropriate to expect lower performing

students to grow by a greater amount than higher performing students on the assessments used for SLOs in this state, growth targets may be set slightly too high for students with lower prescores. This is specifically true of individual and tiered targets, which tend to require greater absolute growth for lower performing students relative to higher performing students.

While I did not find a relationship between teacher SLO scores and classroom percentage of ELLs, there was a small but negative relationship, between the percentage of students with disabilities in a classroom and a teacher's SLO score, -0.22 ($p<.001$).[53] In other words, for every 1 percentage point increase in students with disabilities in a teacher's classroom, a teacher's SLO score *decreases* by approximately 0.22 percentage points. Based on a simple multivariate OLS regression, whereby standardized postscores are regressed on standardized prescores and an indicator variable for whether one is classified as having a learning disability (*swd*), I find that the coefficient on *swd* is negative and significant *(* -0.43). This suggests that controlling for prescore, students with disabilities in the sample do not grow as much from pre to posttest as general education students and therefore may be less likely to reach their growth targets compared to general education students.

Discussion

The purpose of this chapter was to investigate the extent to which teacher-level SLO scores provide an accurate and consistent measure of teacher performance. To do so, I examined how teacher scores were impacted by the choice of student growth target, by comparing teachers scores based on district-set targets to those based on alternative

[53] This correlation is slightly less, -0.20 ($p<.001$), when the sample is limited to teachers in courses with adequate assessment data quality (n=3,324).

student growth targets. I also examined the extent to which teacher SLO scores correlated across courses, across years, and with MGP scores, and examined how the results shifted when the sample was limited to teachers who taught in courses with adequate assessment quality. Finally, I examined the extent to which classroom makeup was related to teacher SLO scores.

The analysis on the sensitivity of teacher performance to student growth targets reveals findings important for assessing the comparability of teacher scores across the state. First, different target formulations will classify teachers differently, based on the distribution of student scores within each classroom. While most teachers will be ranked similarly based on teacher-level scores calculated from alternative target types, a full 25% of teachers would move into a different quartile if the target type were to change from an indidvual to a categorical target, or vice versa.

Second, applying student targets uniformly to every student in the sample in both years reveals that districts vary in the standards they hold students to through the student growth target, particularly in year one, and that changes districts made to the targets from year one to year two substantially altered teacher scores, causing teacher scores to look very different when the growth target is held constant over time. That being said, variation in targets across districts or within-district modifications to the targets may be necessary in order to ensure that student growth targets are realistic given the ability of students and the difficulty of the assessment administered. Ultimately, states and districts need to be attuned to how the relationship between the difficulty of the assessments and difficulty of student targets affects teacher-level scores, given that altering the assessment

or the student growth target can have large consequences on the percentage of students meeting their target and consequently, on teacher-level scores.

The stability of teacher SLO scores over time among teachers in courses with adequate assessments in each year is quite low, and appears to be masked in the full sample due to bias from low quality assessments. The low year-to-year correlation in the reduced sample is likely the result of changes districts made to the assessments and targets which altered the rankings of teachers from one year to the next, evident by the fact that the inter-temporal stability of SLO scores in the reduced sample shifts upward when every teacher score is derived from the same student growth target formula. Therefore, ensuring that the student growth targets and assessments remain stable prior to the inclusion of teacher scores for high-stakes decisions will be necessary for reducing fluctuations in teacher scores and making meaningful year-to-year comparisons.

Within-teacher scores across courses, on the other hand, are fairly stable, particularly for teachers in courses with adequate assessment properties. In fact, the across-course correlations appears to improve from 2012-13 to 2013-14 in the reduced sample of teachers in courses with adequate assessments (from 0.58 to 0.65). While these results provide initial evidence that SLOs scores can provide a stable measure of teacher effectiveness, the correlations are likely influenced by within-district consistency in the implementation of SLOs as well as within-teacher consistency in how SLOs are being administered and graded. Nonetheless, results from analyses on the stability of SLO scores suggest that when based on adequate and consistent assessments and target formulations, SLO scores can provide a fairly stable measure of teacher effectiveness.

Somewhat troubling is the lack of relationship between SLO scores and MGP scores among all teachers in the sample with both metrics, given that both scores are intended to provide a measure of a teacher's contribution to student growth. If the sample in both years were to generalize to the full sample of teachers with SLO and MGP scores, the low SLO-MGP correlation means that 11% of teachers will be considered highly effective under one metric and highly ineffective under the other. This will pose a policy dilemma for the state when both scores are included in final teacher evaluation scores for upwards of 20% of teachers.

In year one, removing assessments with poor quality increases the strength of the relationship between teacher SLO scores and MGP scores, indicating that poor assessment data quality, as identified by anomalous score distributions and low pre-postscore correlations, is biasing the SLO-MGP correlation downward in that year. That being said, the SLO-MGP correlation for the reduced sample of teachers in both years remains low (0.13), particularly which compared to the across-course reliability estimates of SLO scores in both years (0.58 and 0.65, respectively), which provides an upper bound on the SLO-MGP relationship.

It is unclear what SLO-MGP correlation is reasonable to expect. Findings from the MET project, which examined the relationship between within-teacher value-added scores based on different classrooms and different standardized assessments (with different stakes attached to each assessment) found correlations between 0.18 and 0.38 (Bill & Melinda Gates Foundation, 2010). The correlations presented here, based on the reduced sample of teachers with adequate assessment data quality, are on the low end of those found in the MET study. There are several potential reasons for this. First,

variation in classroom makeup and teacher expertise across courses may be driving the small relationship. Second, SLOs may simply be a fairly noisy measure of teacher ability in first two years of implementation, due to differences in how SLOs are created from district to district as well as from course to course. Third, it may be that SLOs and MGPs are actually measuring two different aspects of teacher effectiveness. Therefore, additional research is needed to examine this relationship when controlling for differences in classroom makeup and when the SLO system has stabilized after several years of implementation.

Finally, teachers of students with lower average performance at baseline and teachers with a greater number of SWD students had slightly lower SLO scores in 2013-14, regardless of the inclusion of teacher scores from low-quality assessments. This could be a function of the amount of growth required of these student populations through the student growth targets. On the other hand, teachers in the sample may, in fact, be less effective when teaching students who face certain challenges. A final possibility is that more effective teachers are sorting into higher-performing classrooms, suggesting that this correlation is capturing true differences in teacher effectiveness. This issue requires careful consideration on the part of districts in order to continue to set high expectations for all student subgroups but not place teachers who work with a greater number of at-risk students at a disadvantage in the evaluation system.

Tables and Figures

Table 11. Matched SLO and MGP courses for teachers in the sample.

SLO courses	MGP courses
• 3rd Grade ELA	• 4th or 5th grade ELA
• 3rd Grade Mathematics	• 4th or 5th grade Mathematics
• British Literature • Literature and Composition • World Literature	• Ninth Grade Literature and Composition • American Literature and Composition
• Chemistry I	• Biology • Physics
• Algebra	• Coordinate Algebra • Analytic Geometry
• American Government • World History	• United States History • Economics/Business/Free Enterprise

Table 12. Frequency of teachers in 2012-13 sample (first row) and 2013-14 sample (second row), by district/course.

District	Year	ELA Grade 3	Math Grade 3	British Literature	Literature and Composition	World Literature	Chemistry	Algebra	Am Gov't	World History	Total
District A	12-13		131	25	38		17	20	27	24	282
	13-14		107	15	20		15	17	17	21	212
District B	12-13	152	170		47		26			39	434
	13-14	157	178	35	52		25		21	47	515
District C	12-13	144	152	22		41	22	37	42	41	501
	13-14	140	178	39		66	29	50	48	56	606
District D	12-13	280	287	72		87	70	88	62	71	1,017
	13-14	400	449	90		91	98	97	105	110	1,440
District E	12-13	34		12							46
	13-14	282		40	68		117	58	24	84	673
District F	12-13	124	105	16		42	11	28	56	11	393
	13-14	27	100				8	15	14	18	182
District G	12-13	118	133	27	37		20	34	10	43	422
	13-14	179	152	29	57		19	42	28	47	553
District H	12-13	116	118	28	28		16	28	32	32	398
	13-14	130	124	33	33	5	17		38	27	407

Notes: The first row of each district is the number of teachers in 2012-13 included in the sample, the second row of each district is the number of teachers in 2013-14 included in the sample. Note that only students who had non-missing prescores and postscores, were from district/courses in which the number of students with assessment scores was equal to or greater than 200, and did not have postscores greater than the maximum number of points on the test were used to calculate teacher-level scores.

Table 13. Average teacher scores (percentage of teachers' students meeting their targets), by district and year

	2012-13	2013-14
District A	41.55	68.78
District B	63.82	61.84
District C	10.68	22.36
District D	25.53	39.21
District E	48.96	33.99
District F	54.03	72.35
District G	60.16	66.39
District H	14.92	62.80
N	3,493	4,588

Table 14. Average 2012-13 teacher scores (percentage of teachers' students meeting their target), by district, under each alternative target type (n=3,493)

	Alternative Targets				Difference Between Categorical Target and Individual Target
	Individual Target	Tiered Target	Uniform Target	Categorical Target	
District A	31.8	30.8	30.4	33.7	1.9
District B	41.8	39.2	34.5	49.2	7.4
District C	33.2	34.7	37.2	32.6	-0.6
District D	62.4	65.7	69.0	56.0	-6.4
District E	55.5	54.7	49.3	56.3	0.8
District F	67.7	66.8	66.6	70.8	3.1
District G	47.0	40.7	33.5	39.9	-7.1
District H	62.5	61.6	60.4	63.0	0.5
Total	51.8	51.6	51.2	50.5	0.0

Note. Total average teacher scores under each alternative target type are slightly less than 52% due to the exclusion of certain students and teachers from teacher-level scores as discussed in the Data section. The average difference between the categorical and individual target is based on weighting each district equally.

Table 15. Spearmen rank correlations of teacher scores (percentage of teachers' students meeting their targets) under each alternative target type (n=3,493).

	Individual Target	Tiered Target	Uniform Target	Categorical Target
Individual Target	1.000			
Tiered Target	0.970	1.000		
Uniform Target	0.880	0.929	1.000	
Categorical Target	0.899	0.862	0.820	1.000

Table 16. Percentage of teachers whose score falls in each quartile bin, based on the alternative categorical target and alternative individual target (n=3,493).

		Teacher Scores based on Categorical Target				
		1	2	3	4	Total
Teacher Scores based on Individual target	1	21.56	3.49	0.31	0.14	25.51
	2	2.83	16.58	4.98	0.14	24.53
	3	0.4	4.15	16.81	3.69	25.05
	4	0.57	0.43	2.92	20.98	24.91
	Total	25.37	24.65	25.02	24.96	100

Table 17. Target type, by district.

	2012-13	2013-14
District A	Individual (7)	Individual (7)
District B	Individual (3), Category (2)	Individual (7)
District C	Individual (8)	Individual (8)
District D	Individual (8)	Individual (6), Uniform (2)
District E	Tier (2)	Tier (5), Individual (2)
District F	Individual (6), Uniform (2)	Individual (4), Uniform (2)
District G	Categorical (8)	Categorical (8)
District H	Individual (8)	Individual (8)
Total	Individual (40), Categorical (10), Tier (2), Uniform (2)	Individual (42), Categorical (8), Tier (5), Uniform (4)

Notes. Numbers in parentheses are the number of courses with the given target type. Only courses in the sample with student n-sizes greater than 200 are included.

Table 18. Comparison of district average teacher scores under district-set target and an alternative individual target, by year.

	2012-13			2013-14		
	District Target	Individual Target	Difference between columns 1 and 2	District Target	Individual Target	Difference between columns 4 and 5
District A	41.55	31.79	-9.76	68.78	33.02	-35.76
District B	63.82	41.84	-21.98	61.84	68.37	+6.53
District C	10.68	33.17	+22.49	22.36	36.92	+14.56
District D	25.53	62.39	+36.86	39.21	24.73	-14.48
District E	48.96	55.46	+6.5	33.99	57.65	+23.66
District F	54.03	67.72	+13.69	72.35	77.54	+5.19
District G	60.16	46.97	-13.19	66.39	56.38	-10.01
District H	14.92	62.47	+47.55	62.80	67.14	+4.34
Total	35.94	51.83	10.27	49.57	46.37	-0.75

Notes. N=3,493 teachers in 2012-13; n=4,588 teachers in 2013-14. The individual target is based on the alternative individual growth target requiring 23% growth from the prescore to maximum number of points on the test. The total difference in average performance expressed in the last row of column 3 and 6 are based on weighting the average difference in performance between district-set targets and the individual target from each district equally.

Table 19. Comparison of change in district average teacher scores under district-set target and an alternative individual target.

	District-set Target			Alternative Individual Target		
	2012-13	2013-14	Change over time	2012-13	2013-14	Change over time
District A	41.55	68.78	27.31	31.79	33.02	1.23
District B	63.82	61.84	-1.98	41.84	68.37	26.53
District C	10.68	22.36	11.68	33.17	36.92	3.75
District D	25.53	39.21	13.68	62.39	24.73	-37.66
District E	48.96	33.99	-14.97	55.46	57.65	2.19
District F	54.03	72.35	18.32	67.72	77.54	9.82
District G	60.16	66.39	6.23	46.97	56.38	9.41
District H	14.92	62.80	47.88	62.47	67.14	4.67

Notes. N=3,493 teachers in 2012-13; n=4,588 teachers in 2013-14. The alternative target is based on an individual growth target requiring 23% growth from the prescore to maximum number of points on the test.

Table 20. Percentage of teachers in each quartile bin of the SLO and MGP score distribution (n=794).

		MGP Quartiles				
SLO Quartiles		1	2	3	4	Total
	1	12.34	9.70	9.82	6.05	37.91
	2	1.89	3.15	3.90	3.15	12.09
	3	6.05	5.92	6.17	6.93	25.06
	4	4.79	6.17	5.16	8.82	24.94
	Total	25.06	24.94	25.06	24.94	100.0

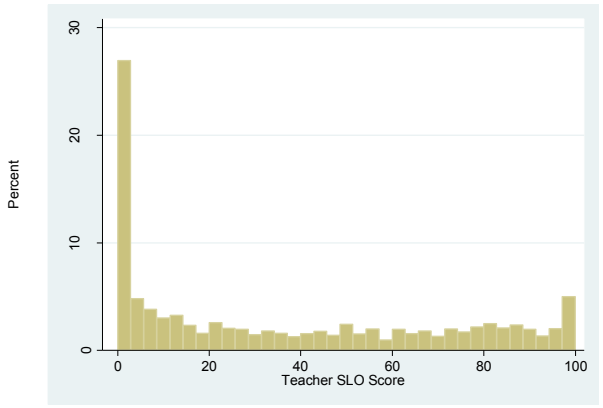
Note. The total percentage of teachers in the first SLO quartile (row 1) is greater than 25 (and the total percentage of teachers in the second SLO quartile (row 2) is less than 25) due to the high number of teachers with an SLO score of 0 that are grouped into the first quartile.

Table 21. Within-teacher across-course 2012-13 SLO-MGP correlation by MGP tercile (n=794).

	Correlation	P-value	Average SLO Score	N
MGP<=33	-0.11	0.25	17.8	105
33<MGP<=66	0.12	0.00	24.0	636
MGP>66	0.24	0.07	28.8	53

Figure 7. The frequency of teacher SLO scores in the 2012-13 sample (Panel A) and 2013-14 sample (Panel B).

Panel A. 2012-13



Panel B. 2013-14

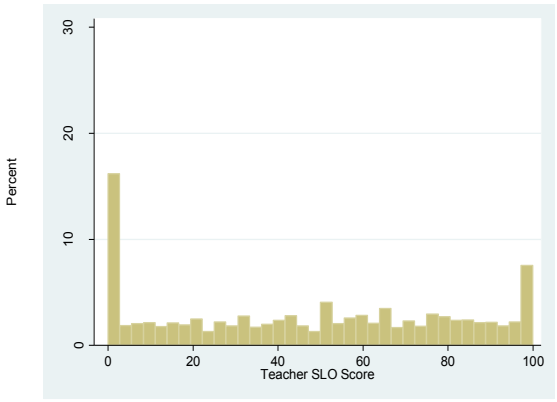
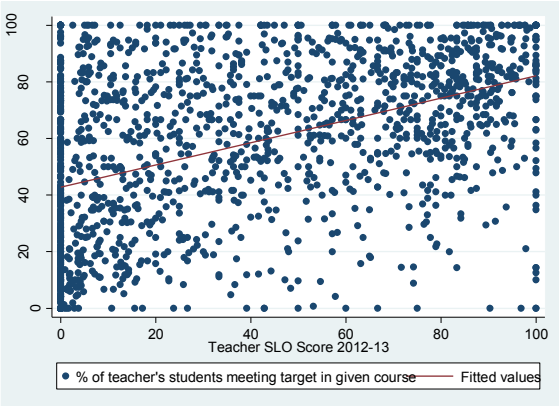


Figure 8. Scatter plot of teacher SLO scores in the 2013-14 sample (y-axis) and 2012-13 sample (x-axis), in the full sample of teachers (Panel A) and the reduced sample of teachers with adequate assessments (Panel B), with a linear best fit overlaid.

Panel A. Full Sample (n=1,593)



Panel B. Reduced Sample (n=595)

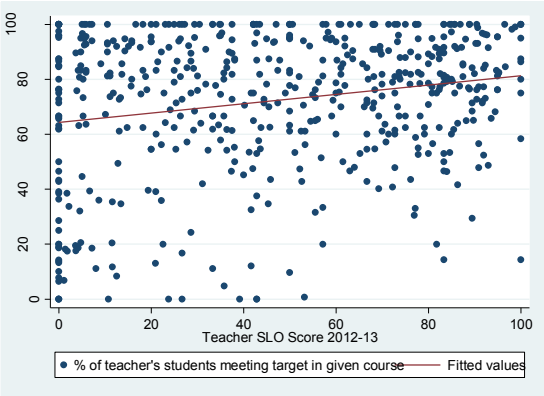
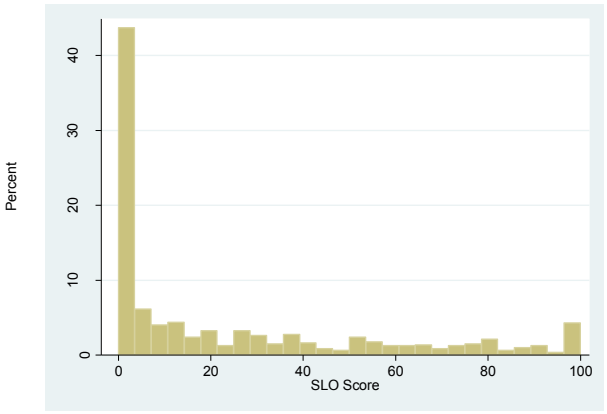


Figure 9. Distributions of teacher SLO scores (Panel A) and MGP scores (Panel B) for teachers in the SLO-MGP sample (n=794).

Panel A: SLO scores



Panel B: MGP scores

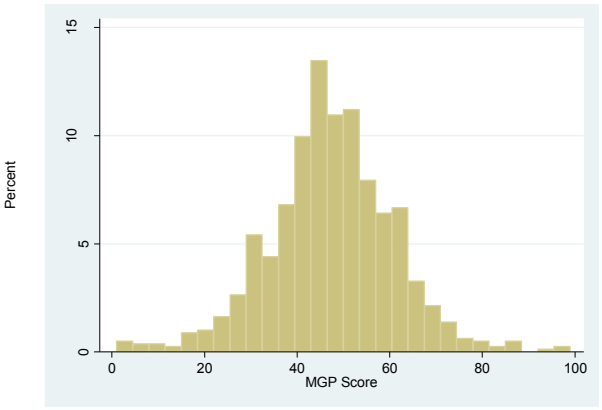


Figure 10. Distribution of 2012-13 average student implied gains by teacher and course, for teachers in the SLO-MGP sample.

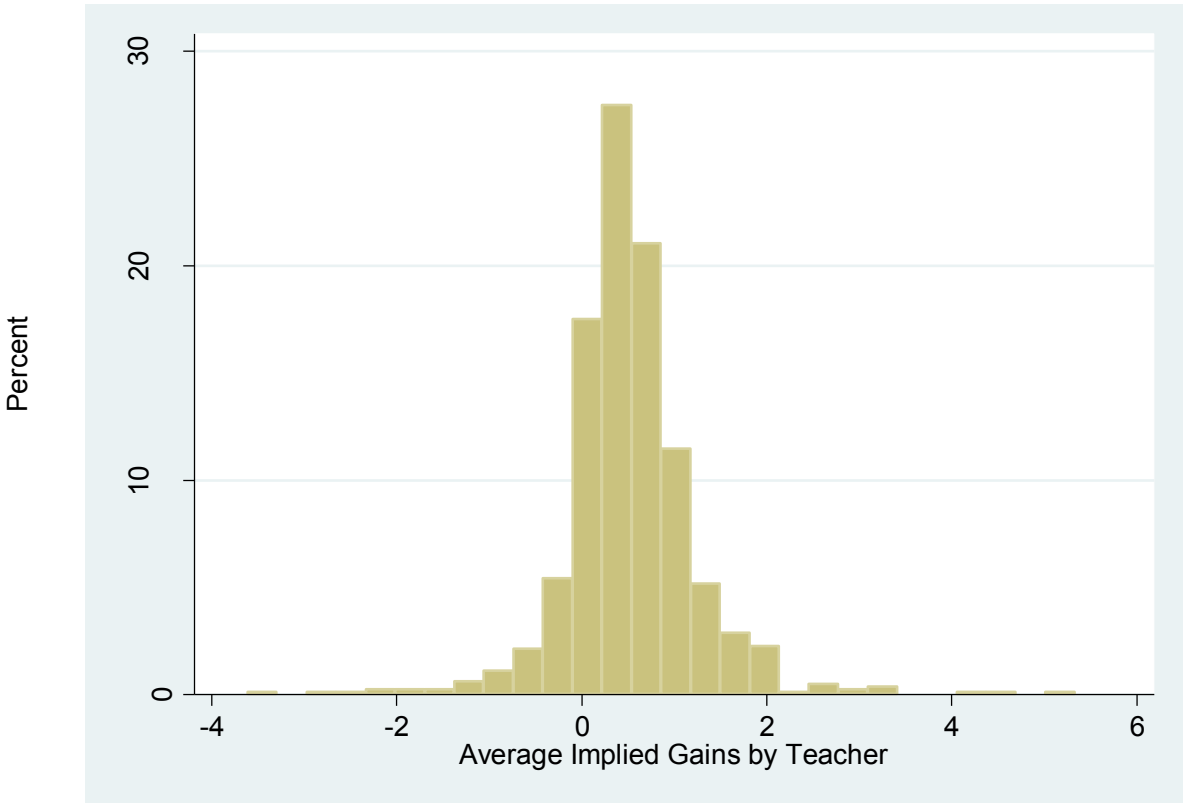
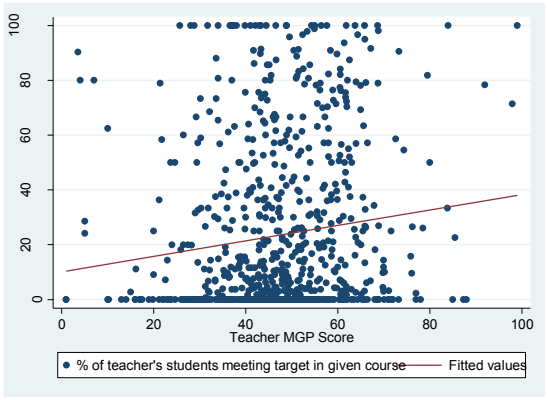
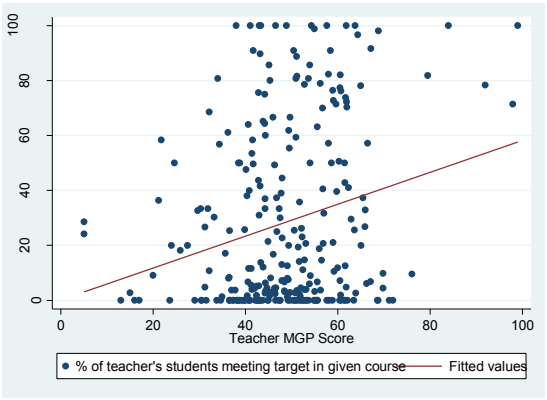


Figure 11. Scatter plot of teacher 2012-13 SLO scores (y-axis) and teacher 2012-13 MGP scores (x-axis), in the full sample of teachers (panel A) and reduced sample of teachers with adequate assessments (panel B) with linear best fit overlaid.

Panel A. Full sample (n=794)



Panel B. Reduced Sample (n=268)



Dissertation Conclusion

Creating an SLO that provides a valid metric of student and teacher ability is a difficult enterprise, even more so is ensuring that the SLO metric demands the same standard across classrooms, courses and districts (Lachlan-Hache et al., 2013; Marion et al., 2012). The findings presented in both chapters suggest that the quality of the assessment and the type and standard of the student growth target will impact student and teacher-level scores. I posit that issues related to the assessments and growth targets can be addressed through (a) more targeted resources around the development and assessment of the validity of tests used in a high stakes framework (see Herman et al., 2011); (b) state guidance on appropriate standard-setting procedures for determining ambitious yet realistic student growth targets; and (c) standardized procedures and guidelines regarding the administration of assessments, scoring of assessments, and calculation of student target scores.

While results from the analyses illuminate several problems with the quality of student SLO data being produced, the purpose of these studies is not to advocate for an increase in the standardization of tests or growth targets, since doing so may lead to less alignment between the SLO and the curriculum and student body makeup of each classroom, and remove any instructional benefits of the SLO. Moreover, given time and financial constraints, it is impossible for states to create standardized assessments for every course/grade with score reliability at the level of standardized assessments (Steele et al., 2011). Instead, the results presented in Chapters 1 and 2 suggest the need for greater investigation into how differences in the quality of assessments and choice of targets are affecting student and teacher scores well in advance of the scores being used

for high stakes decisions. This is particularly true since these findings provide suggestive evidence that changes that districts made to the SLO components between year one and year two of implementation improved the quality of assessment data, the comparability of student and teacher scores, and the stability of teacher scores across courses.

Ultimately, this dissertation provides a framework for states and districts seeking to evaluate the validity and reliability of inferences from student and teacher-level SLO scores based on the quality of assessment data and the choice of student growth targets. The trade-offs in design around standardization versus curriculum alignment illuminated by findings presented in Chapters 1 and 2 are crucial for policymakers to consider when designing or modifying their SLO system.

References

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*, 95–135.

American Educational Research Association/American Psychological Association/National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Institutes for Research *Selecting measures. Database on state teacher evaluation policies*. Center on Great Teachers & Leaders. Retrieved from: http://resource.tqsource.org/stateevaldb/Compare50States.aspx

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37–65.

Betebenner, D. W. (2009). *Growth, standards and accountability*. Dover, NH: The Center for Assessment. Retrieved from http://www.nciea.org/publication_PDFs/growthandStandard_DB09.pdf

Betebenner, D. (2014, December 29). Personal email communication.

Bill and Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study* (Final research report). Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

Bill & Melinda gates Foundation. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project*. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf.

Campbell, D. T. (1975). Assessing the impact of planned social change. G. M Lyons, ed, *Social Research and Public Policies: The Dartmouth/OECD Conference*. Hanover, NH: Public Affairs Center, Dartmouth College.

Castellano, K. E., and Ho, A. D. (2013). *A practitioner's guide to growth models*. Council of Chief State School Officers, Washington DC.

Cheng, M.Y., Fan, J., Marron, J.S., 1993. Minimax efficiency of local polynomial fit estimators at boundaries. Unpublished manuscript Series # 2098, Institute for Statistics, University of North Carolina.

Cheng, M., Y. (1994). On boundary effects of smooth curve estimators (dissertation).

Unpublished manuscript Series # 2319, Institute for Statistics, University of North Carolina.

Chetty, R., Friedman J. F., and Rockoff, J. E. (2011). Long-Term Impact of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. Washington DC: National Bureau of Economic Research, Working Paper 17699.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2007b). Teacher-student matching and the assessment ofteacher effectiveness. *Journal of Human Resources*, *41*(4), 778–820.

Cole, R., Haimson, J., Perez-Johnson, I., and May, H. (2011). *Variability in pretest-posttest correlation coefficients by student achievement level*. NCEE Reference Report 2011-4033. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Dee, T. S., Dobbie, W., Jacob, B., McCrary, J. and Rockoff, J. (2011). *Manipulation on the NYS regents exam.* Working paper. Retrieved from www.personal.umich.edu/~bajacob/files/Miscellaneous/manipulation-of-grading-ny-reg-exams.pdf

Diaz-Bilello, E. K. (2011). *A validity study of interim assessments in an urban school district*. (Unpublished doctoral dissertation). University of Colorado, Boulder.

Gitomer, D. H. & Bell, C. A. (in press). Evaluating teaching. In K. F. Geisinger (Ed.) *APA Handbook of Testing and Assessment in Psychology*.

Glazerman, S., Goldhaber, D., Loeb S., Raudenbush, S., Staiger, D., & Whitehurst, G.J. (2011). Passing muster: Evaluating teacher evaluation systems. Washington, DC: The Brookings Institution.

Goe and Holdheide. (2011). *Measuring teachers contribution to student learning growth for the other 69%.* National Comprehensive Center for Teacher Quality.

Goldhaber, D. (2006). Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources*, *42*(4), 765–794.

Goldhaber, D and Hansen, M. (2012). *Is it just bad class? Assessing the long-term stability of estimated teacher performance*. Washington DC: American Institutes for Research, CALDER, Working Paper 73.

Goldhaber D., and Theobald, R *Do different value-added models tell us the same things?* Carnegie Foundation for the Advancement of Teaching, Knowledge Brief 4. Retrieved from http://carnegieknowledgenetwork.org/briefs/value-added/different-growth-models/.

Goldschmidt, P., Choi, K., and Beaudoin, J.P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance*. Washington, DC: Council of Chief State School Officers.

Harris, D. N. (2012). *How do value-added indicators compare to other measures of teacher effectiveness?* Carnegie Knowledge Network, Issue Brief 5.

Harris, D. N., and Sass, T. R. (2008). *Teacher training, teacher quality, and student achievement* [Working Paper No. 3]. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Herman, J. L., Heritage, M., and Goldschmidt, P. (2011). *Developing and selecting assessments of student growth for use in teacher evaluation systems (extended version).* Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ho, A. D. and Yu, C. C. (2014). Descriptive Statistics for Modern Test Score Distributions: Skewness, Kurtosis, Discreteness, and Ceiling Effects. *Educational and Psychological Measurement*, 1-24.

Hoffer, T. B., Hedberg, E. C., Brown, K. L., Halverson, M. L., Reid-Brossard, P., Ho, A. D., and Furgol, K. (2011). *Final report on the evaluation of the growth model pilot project*. Washington DC: US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

Hill, H. C., Kapitula, L. & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, 48 (3), 794-831.

Jacob, B. and Levitt, S. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics*. 118(3): 843-877.

Kane, T. and Staiger, D. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation.* National Bureau of Economic Research, Working Paper, 14607. Retrieved from http://www.nber.org/papers/w14607

Kane, M. (2006). Validity in Educational Assessments. In Brennan, R. L., (ed) *Educational Measurement, 4th Edition*. Portsmouth NH: Praeger

Koedel, C. & Betts, J. R. (2010). Value-added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy 5*, 54-81.

Koretz, D. (2008). *Measuring up: What educational testing really tell us*. Cambridge, MA: Harvard University Press.

Koretz, D. M., and Barron, S. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)* [No. MR-1014-EDU]. Santa Monica, CA: RAND Corporation. Retrieved November 15, 2014 from http://www.rand.org/pubs/monograph_reports/MR1014/

Koretz, D. M., and Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (Fourth ed., pp. 531–578). Westport, CT: Praeger.

Kupermintz, H. Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation and Policy Analysis*, 25, 287–298.

Lachlan-Haché, L., Cushing, E., & Bivona, L. (2012). *Student learning objectives as measures of educator effectiveness: The basics.* Washington, DC: American Institutes for Research. Retrieved from http://educatortalent.org/inc/docs/SLOs_Measures_of_Educator_Effectiveness.pdf

Lachlan-Haché, L.,Matlach, L., Reese, K., Cushing, E., Mean, M. (2013). *Student learning objectives: Early lessons learned from the Teacher Incentive Fund*. Washington, DC: American Institutes for Research.

Lacireno-Paquet, N., Morgan, C., & Mello, D. (2014). *How states use student learning objectives in teacher evaluation systems: a review of state websites* (REL 2014–013). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory North-east & Islands. Retrieved from http://ies.ed.gov/ncee/edlabs.

Lamb, L. M., Schmitt, L. N. T., and Cornetto, K. M. (2011). *AISD REACH program update: 2009-2010 TAKS school-wide growth* (DRE Publication No. 09.81). Austin, TX: Austin Independent School District.

Liguori, B. J. (2011). *High stakes testing and teacher resistance: New York City schools in an era of increased accountability*. (Unpublished doctoral dissertation). Cornell University, New York.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 44(1), 47–67.

Loeb, S. and Candelaria, C. A. (2013). *How stable are value-added estimates across years, subject and student groups*. Carnegie Knowledge Network, Knowledge Brief 3.

Marion, S. F., & Buckley, K. H. (2011). *Approaches and considerations for incorporating student performance results from "non-tested" grades and subjects into educator effectiveness determinations.* Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from http://www.nciea.org/publications/Considerations%20for%20non-tested%20grades_SMKB2011.pdf

Marion, S.F. & Buckley, K. H. (In press). Design and implementation considerations of performance-based and authentic assessments for use in accountability systems. In Braun, H. (ed). *Meeting the Challenges to Measurement in an Era of Accountability*. Washington, DC: NCME.

Marion, S., DePascale, C., Domaleski, C., Gong, B., Diaz-Bilello, E. (2012). *Considerations for analyzing educators' contributions to student learning in non-tested subjects and grades with a focus on student learning objective*s. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from http://www.nciea.org/publication_PDFs/Measurement%20Considerations%20for%20NTSG_052212.pdf

May, H., Perez-Johnson, I., Haimson, J., Sattar, S., and Gleason, P. (2009). *Using State Tests in Education Experiments: A Discussion of the Issues* (NCEE 2009-013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

McCaffrey, D. F., Koretz, D., Lockwood, J. R., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis,A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67–101.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education, Finance and Policy*, 4(4), 572–606.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142, 698–714

Papay, J. (2011). Different tests, different answers: The stability of teacher value added estimates across outcome measures. *American Educational Research Journal*, 48(1): 163-193.

Pianta, R. C., Hamre, B. K., Haynes, N. J, Mintz, S., La Paro, K. M. (2006). *CLASS classroom assessment scoring system: Manual middle secondary version pilot*.

Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2009). *The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades.* Washington, DC: Center for Educator Compensation Reform. Retrieved from http://cecr.ed.gov/pdfs/guide/other69Percent.pdf

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2), 247–252.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville: University of Tennessee, Value-Added Research and Assessment Center.

Sass, T. R. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. National Center for Analysis of Longitudinal Data in Education Research Brief 4. Retrieved from: http://www.urban.org/UploadedPDF/1001266_stabilityofvalue.pdf.

Schmitt, L. N. T. (2011). *AISD REACH program update, 2010-2011: Texas Assessment of Knowledge and Skills growth and student learning objectives* (DRE Publication No. 10.84 RB). Austin, TX: Austin Independent School District

Schmitt, L. N. T., Cornetto, K. M., Lamb, L. M., & Imes, A. (2009). *AISD REACH year 2 evaluation report I, 2008-2009* (DRE Publication No. 08.53). Austin, TX: Austin Independent School District.

Schmitt, L. N. T., and N. Ibanez. (2011). *AISD REACH program update: 2009-2010 Texas Assessment of Knowledge and Skills (TAKS) results and student learning objectives (SLOs).* Austin, TX: Austin Independent School District Department of Program Evaluation.

Schmitt, L. N. T., Lamb, L. M., Cornetto, K. M., & Courtemanche, M. (2013). *AISD REACH program update, 2012-2013: Student learning objectives* (DRE Publication No. 12.83a). Austin, TX: Austin Independent School District

Schmitt, L. N. T., Lamb, L. M., Cornetto, K. M., & Courtemanche, M. (2014). *AISD REACH program update, 2012-2013: Student learning objectives* (DRE Publication No. 12.83b). Austin, TX: Austin Independent School District.

Scott, D. (2013). *Will Common Core Help or Hurt Schools' Cheating Problem?* Washington DC: Governing the States and Localities. Retrieved November, 20, 2014 from http://www.governing.com/topics/education/gov-will-common-core-help-hurt-cheating.html

Shepard, L. (2012). *Evaluating the use of tests to measure teacher effectiveness: Validity as a theory-of-action framework*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Slotnick, W. J., Smith, M. D., Helms, B. J. and Quao, Z. (2004). *Catalyst for Change: Pay for Performance in Denver Final Report.* Boston, MA: Community Training and Assistance Center.

Slotnick, W. J., Smith, M. D., Helms, B. J. and Quao, Z. (2013). *It's more than money: Teacher Incentive Fund – Leadership for Educators' Advanced Performance Charlotte-Mecklenburg Schools* Boston, MA: Community Training and Assistance Center.

Steele, J. L., Hamilton, L. S., Stecher, B. M. (2011). *Incorporating student performance measures into teacher evaluations systems*. Santa Monica, Ca: RAND.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., Mario, A., Fisher, T., and Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, 29(1), 11–35.

The Reform Support Network (2011). *Targeting Growth Using Student Learning Objectives as a Measure of Educator Effectiveness*. US Department of Education. Retrieved from http://msde.state.md.us/tpe/TargetingGrowth_Using_SLO_MEE.pdf

The Reform Support Network. (2014). *A Toolkit for Implementing High-Quality Student Learning Objectives 2.0*. US Department of Education. Retrieved from http://www.engageny.org/sites/default/files/resource/attachments/rsn-slo-toolkit.pdf

Tyler, J. (2011). *Designing High Quality Evaluation Systems for High School Teachers: Challenges and Potential Solutions*. Center for American Progress. Retrieved from http://www.americanprogress.org/issues/education/report/2011/11/29/10614/designing-high-quality-evaluation-systems-for-high-school-teachers/

US Department of Education. (2010). *Race to the Top program guidance and frequently asked questions*. Washington DC. Retrieved from http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCcQFjAA&url=http%3A%2F%2Fwww2.ed.gov%2Fprograms%2Fracetothetop%2Ffaq.pdf&ei=LONGU-uxKpLNsQS29YLwAw&usg=AFQjCNEBoTn4J0K8whUrLnc8Rwigw1LZaA&sig2=uY8_nhn78l1NgQkVTgSASg&bvm=bv.64507335,d.cWc&cad=rja

Vigdor, J. L. (2008). Scrap the sacrosanct salary schedule. *Education Next*, *8*(4), 36–42.

Weisberg, D., Sexton, S., Mulhern, J., and Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: New Teacher Project.

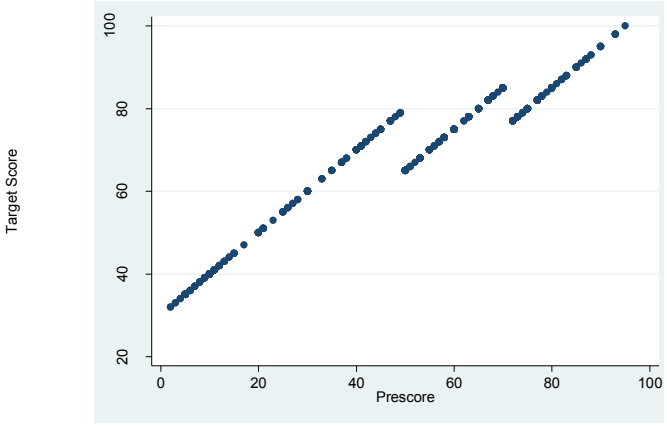
Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, 1(1), 57–67.

Appendix A

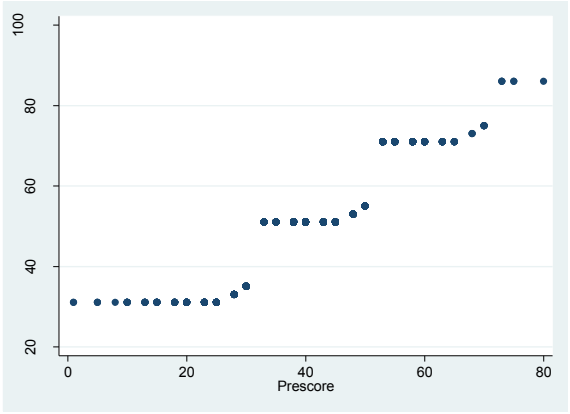
Unintended Consequences of Target Type. Certain target types may have unintended consequences. For example, the tiered and categorical targets, by the nature of their target formulations, exacerbate discontinuity in the prescores around the cut-points. In Figure A1 panel A, I display a scatter plot of each student's prescore and target score for District E in Mathematics Grade 3. Evident is that students who score on either side of the border of a cut-point have drastically different target scores; for example, a student who receives a 49 on the pretest is required to score 15 points higher than a student who receives a 50. A similar but opposite issue occurs with categorical targets, whereby a student on the right of the cut-score has substantially higher target score than a student to the left of the cut-score, as with student Chemistry scores in District G (Panel B).

Figure A1. Scatter plot of student target scores by prescores in two district/courses, illustrating discontinuities at the cut-points.

Panel A. District E, Tiered Target



Panel B. District G Categorical Target



Note: The target formulation for District G's categorical targets contains the additional requirement that students scoring within three points of their tier's ceiling must increase their post-assessment scores by at least five points in order to be considered meeting their growth targets

Students with Missing Prescores. An additional issue with calculating target scores is how to determine SLO attainment for students who fail to take the pre-assessment. In one district in this state, students with missing prescores were assigned the lowest possible target score. This can introduce an upward bias into student SLO attainment since it is unlikely that every student with a missing prescore would have scored the lowest possible number of points on the pretest.

Appendix B

Table B1. Number of students, by district and course, included in the analysis of score distortion in 2012-13.

	ELA third grade	Math third grade	British Literature	Literature and Comp	Chemistry	Algebra	American Gov't	World History	Total
District A	0	322	721	479	1,491	273	530	919	4,735
District B	2,364	0	1,819	1,328	2,622	0	0	1,668	9,801
District G	1,114	901	978	577	1,106	732	321	1,024	6,753
District F	0	402	559	0	0	0	0	0	961
Total	3,478	1,625	4,077	2,384	5,219	1,005	851	3611	22,250

Note. A district/course cell with a zero does not indicate the absence of students; rather that the given cell did not meet the criteria required (e.g., a teacher teaches the same course in both years).

Table B2. Number of students, by district and course, included in the analysis of score distortion in 2013-14.

	ELA third grade	Math third grade	British Literature	Literature and Comp	Chemistry	Algebra	American Gov't	World History	Total
District A	0	263	789	539	1,456	331	601	952	4,931
District B	2,309	0	2,178	1,729	2,412	0	0	1,725	10,353
District G	1,294	1,104	1,399	813	1,188	927	557	1,324	8,606
District F	0	740	952	0	0	0	0	0	1,692
Total	3,603	2,107	5,318	3,081	5,056	1,258	1,158	4,001	25,582

Note. A district/course cell with a zero does not indicate the absence of students; rather that the given cell did not meet the criteria required (e.g., a teacher teaches the same course in both years).

Appendix C

Table C1: Frequency of 2012-13 student data excluded from teacher-level scores, by district and course.

	ELA grade 3	Brit Lit	Lit and Comp	World Lit	Chemistry	Math grade 3	Algebra	American Gov't	World History	Total
District A	0	0	0	0	1	1	0	0	0	2
District B	0	0	0	0	0	23	0	0	0	23
District C	0	0	0	0	2	0	0	0	0	2
District D	0	0	0	0	0	0	0	0	0	0
District E	0	0	0	0	0	0	0	0	0	0
District H	208	3	0	50	43	90	30	158	10	592
District G	0	0	0	0	0	0	0	0	0	0
District H	0	0	0	0	0	0	0	0	0	0
Total	208	3	0	50	46	114	30	158	10	619

Note that the majority of 2012-13 excluded data comes from District F, which had a considerable number of missing student prescores.

Table C2: Frequency of 2013-14 student data excluded from teacher-level scores, by district and course.

	ELA grade 3	Brit Lit	Lit and Comp	World Lit	Chemistry	Math grade 3	Algebra	American Govt	World History	Total
District A	0	0	0	0	3	2	0	0	0	5
District B	0	0	0	0	0	0	0	0	0	0
District C	0	0	0	0	0	0	0	0	0	0
District D	556	133	0	180	240	540	163	440	338	2,590
District E	0	0	0	0	24	0	0	1	45	70
District F	0	0	0	0	0	0	0	0	0	0
District G	0	0	0	0	0	6	0	0	1	7
District H	0	0	0	0	0	0	0	0	0	0
Total	556	133	0	180	267	548	163	441	384	2,672

Note that the majority of 2013-14 excluded data comes from District D, which had a considerable number of missing student prescores.

Vita

Katie Hills Buckley

2000-2004	Providence College Providence, Rhode Island	B.A. May, 2004
2004-2006	Georgetown Public Policy Institute Georgetown University Washington, DC	M.P.P May, 2006
2006-2000	Research Associate, Education Development Center Newton, MA	
2008-2015	Doctor of Education Candidate Graduate School of Education Harvard University	
2009-2010	Teaching Fellow Graduate School of Education Harvard University	
2009-present	Graduate Fellow Center for Education Policy Research Harvard University	
2010-present	Consultant The Center for Assessment Dover, NH	
2014	Graduate School of Education Harvard University	M. Ed. November, 2014